Statistical Methods for studying Genetic Variation in Populations

Suyash Shringarpure

August 2012
CMU-ML-12-105

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Statistical Methods for studying Genetic Variation in Populations	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Carnegie Mellon University,School of Computer Science,Machine Learning Department,Pittsburgh,PA,15213	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The study of genetic variation in populations is of great interest for the study of the evolutionary history of humans and other species. Improvement in sequencing technology has resulted in the availability of many large datasets of genetic data. Computational methods have therefore become quite important in analyzing these data. Two important problems that have been studied using genetic data are population stratification (modeling individual ancestry with respect to ancestral populations) and genetic association (finding genetic polymorphisms that affect a trait). In this thesis, we develop methods to improve our understanding of these two problems. For the population stratification problem, we develop hierarchical Bayesian models that incorporate the evolutionary processes that are known to affect genetic variation. By developing mStruct, we show that modeling more evolutionary processes improves the accuracy of the recovered population structure. We demonstrate how nonparametric Bayesian processes can be used to address the question of choosing the optimal number of ancestral populations that describe the genetic diversity of a given sample of individuals. We also examine how sampling bias in genotyping study design can affect results of population structure analysis and propose a probabilistic framework for modeling and correcting sample selection bias. Genome-wide association studies (GWAS) have vastly improved our understanding of many diseases. However, such studies have failed to uncover much of the variation responsible for a number of common multi-factorial diseases and complex traits. We show how artificial selection experiments on model organisms can be used to better understand the nature of genetic associations. We demonstrate using simulations that using data from artificial selection experiments improves the performance of conventional methods of performing association. We also validate our approach using semi-simulated data from an artificial selection experiment on Drosophila Melanogaster.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**167**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Statistical Methods for studying Genetic Variation in Populations

Suyash Shringarpure

CMU-ML-12-105
August 2012

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Eric Xing (Chair)
Stephen Fienberg
Kathryn Roeder
Martin Kreitman

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by the National Science Foundation under grant numbers CCF0523757, DBI0546594, DBI0640543 and IIS0713379, the National Institute of Health under grant numbers 1R01AG02314101, 1R01GM087694, 1R01GM093156, the Defense Advanced Research Projects Agency under grant number Z931302 and by a Presidential Fellows in the Life Sciences Award.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Genetic variation, population genetics, population structure, ancestry inference, artificial selection, association

To the memory of my grandmother (1932-2005).

Abstract

The study of genetic variation in populations is of great interest for the study of the evolutionary history of humans and other species. Improvement in sequencing technology has resulted in the availability of many large datasets of genetic data. Computational methods have therefore become quite important in analyzing these data. Two important problems that have been studied using genetic data are population stratification (modeling individual ancestry with respect to ancestral populations) and genetic association (finding genetic polymorphisms that affect a trait). In this thesis, we develop methods to improve our understanding of these two problems.

For the population stratification problem, we develop hierarchical Bayesian models that incorporate the evolutionary processes that are known to affect genetic variation. By developing *mStruct*, we show that modeling more evolutionary processes improves the accuracy of the recovered population structure. We demonstrate how nonparametric Bayesian processes can be used to address the question of choosing the optimal number of ancestral populations that describe the genetic diversity of a given sample of individuals. We also examine how sampling bias in genotyping study design can affect results of population structure analysis and propose a probabilistic framework for modeling and correcting sample selection bias.

Genome-wide association studies (GWAS) have vastly improved our understanding of many diseases. However, such studies have failed to uncover much of the variation responsible for a number of common multi-factorial diseases and complex traits. We show how artificial selection experiments on model organisms can be used to better understand the nature of genetic associations. We demonstrate using simulations that using data from artificial selection experiments improves the performance of conventional methods of performing association. We also validate our approach using semi-simulated data from an artificial selection experiment on *Drosophila Melanogaster*.

Acknowledgments

I would like to thank my advisor, Prof. Eric Xing, for his constant guidance and support. His enthusiasm for research and his work ethic have been a great source of inspiration, and I have learned a lot from him. I also thank my committee members, Prof. Stephen Fienberg, Prof. Kathryn Roeder, and Prof. Martin Kreitman, for agreeing to be on my committee and providing many helpful suggestions and inputs about the directions this thesis should explore.

The current and former members of the SAILING lab have been great sources of inspiration and help. All of the work in this thesis is much stronger thanks to their feedback and suggestions. They have also made my time at CMU much more enjoyable. Special thanks to Hetunandan Kamisetty, Seunghak Lee, Kriti Puniyani, Pradipta Ray, and Kyung-Ah Sohn for their support.

Thanks to Diane Stidle and Michelle Martin for many years of patient help with countless queries - from mundane queries like "How do I find this room?" to important ones like "What do I need to do to graduate?". Their help with all administrative matters has been invaluable.

My time in graduate school has been a lot of fun thanks to many amazing friends. While there are too many to list here, a few names need special mention: thanks to Apurva Samudra, Swapnil Patil, and Shweta Shah for being such great friends. I also owe thanks to Neeti Wagle and Swati Yelgulwar for their ever-present support. The CMU Quiz club and the Carnegie Library of Pittsburgh have provided much-needed entertainment during my time at CMU.

Prof. Soumen Chakrabarti was my advisor at the Indian Institute of Technology, Bombay. He introduced me to Machine Learning and the fascinating problems therein. His advice was instrumental in my decision to choose CMU. For that I am grateful to him.

Lastly, and most importantly, I am extremely grateful to my family. My parents, my brother, and my extended family have always believed in me and have been supportive of whatever I have wanted to do. Without their efforts and sacrifices, this dissertation would have been impossible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Improvements in sequencing technologies, coupled with their decreasing costs, have made available a number of datasets for the study of genetic variation in populations, such as the Human Genome Diversity Project (HGDP) [Cann et al., 2002, Cavalli-Sforza, 2005], HapMap [Gibbs, 2003] and the 1000 Genomes project [Altshuler et al., 2010]. These data have been used to improve our understanding of the evolutionary history of populations by studying population structure [Bowcock et al., 1994, Novembre et al., 2008, Rosenberg et al., 2002, Tang et al., 2005], population expansions, contractions and migrations [Cavalli-Sforza et al., 1994, Hammer et al., 1998, Reich et al., 2009, Templeton, 2002], mutation rates [Kelly et al., 1991, Valdes et al., 1993, Zhivotovsky et al., 2004], linkage and recombination rates [Conrad et al., 2006]. Studies of genetic variation have also been used to find loci associated with diseases [Consortium, 2007, Cordell and Clayton, 2005, Manolio et al., 2009]. Associated loci have been discovered for diabetes [Saxena et al., 2007, Sladek et al., 2007], prostate cancer [Eeles et al., 2008, Thomas et al., 2008], breast cancer [Antoniou et al., 2008, Easton et al., 1993], Crohn's disease [Libioulle et al., 2007]. Hindorff et al. [2009] catalog the results of a number of disease association studies.

This thesis proposes statistical methods to address two important problems in studying genetic variation in populations - (i) detecting population structure and (ii) understanding the nature of genetic associations. For the former problem, previous attempts have been forced to

make simplifying assumptions or limit model complexity in order to develop tractable methods. We develop hierarchical paramatric and non-parametric Bayesian models of population evolution that more accurately reflect reality while allowing efficient inference. For the latter, current approaches have had significant success in some attempts (as listed above) but have made only limited progress in explaining how genotypic variation accounts for phenotypic variation. We propose the use of artificial selection experiments in model organisms combined with spare regression techniques to explore the effect and frequency spectra in which causal variants may lie.

Chapter 2 introduces the preliminaries that provide context for the work presented in this thesis. In chapter 3, we develop a model for population structure that can take into account the evolutionary processes of admixture and mutation that shape genomes in real populations [Shringarpure and Xing, 2009]. Using a hierarchical Bayesian modeling framework, we show how incorporating these processes in our *mStruct* model improves the accuracy of population structure detection and ancestry inference. By analyzing data from the HGDP, we demonstrate how *mStruct* enables us to not only examine not only population structure, but also qualitatively compare the age of populations. Such comparisons can be used to validate hypotheses about human migration across the world.

A frequent question in population structure analysis is to decide the number of ancestral populations that should be used to best capture the genetic variation observed in a given dataset. We propose *StructHDP* in chapter 4 to address this question using a nonparametric Bayesian method [Shringarpure et al., 2011]. *StructHDP* can be used to detect population structure and choose the optimal number of ancestral populations simultaneously, without requiring user input. We show using previously-studied datasets that the number of populations chosen by *StructHDP* agrees with previous analyses. *StructHDP* thus enables users to impose a Bayesian prior on the number of ancestral populations and make model selection a part of the inference process rather than a post-processing step.

Chapter 5 examines sampling bias in genotyping individuals for population genetic studies. We study the problem of sample selection bias and its effect on ancestry inference in population structure analyses. We propose a probabilistic framework in which this problem can be studied and demonstrate that it can have significant effects on population structure analysis using simulated and real data. We also suggest a simple correction that can eliminate the effects of sample selection bias.

Finally, in chapter 6, we propose that artificial selection experiments on *Drosophila Melanogaster* can be used to generate data well-suited for association studies. The resulting data eliminates a number of the issues that commonly occur in association studies and make the association problem hard to examine. We demonstrate how sparse regression methods can be used to effectively solve the resulting association problems for a large range of causal allele frequencies and effect sizes. This presents a way of examining the allele frequency and effect spectrum of causal variants for complex traits, where existing approaches have had limited success.

Chapter 2

Background

The genetic diversity observed within populations is a result of various evolutionary processes that act on populations. A vast number of evolutionary processes act on populations. They include mutation, recombination, admixture, selection, migration and population expansions or contractions. This thesis aims at studying the genetic variation in populations. It is therefore essential to understand the nature of these evolutionary processes which affect genetic variation. In this chapter, we describe the nature of genetic variation we wish to study and the effects various evolutionary processes have on observed genetic variation. We describe specific ways of studying genetic variation at the population level that are of interest for this thesis and introduce some related work on these aspects.

2.1 Describing genomic diversity and evolutionary processes

Diploids organisms like humans have two copies of each chromosome, one inherited from the mother and one inherited from the father. Each chromosomal copy is called the haplotype, and the two are jointly called the genotype. The DNA copying process that directs this inheritance is not perfect and there can be errors during copying DNA from parent to offspring. This process of imperfect copying of DNA is called mutation. A simple example of mutation involves

copying errors at a single nucleotide/location in the genome, suppose a parental chromosome has nucelotide 'T' at a specific location, the imperfect copying could lead to a child having nucleotide 'C' at that location. The polymorphisms that arise from single/point mutations are called 'single-nucleotide polymorphisms" or SNPs. Each variant observed at a SNP is called an allele. These polymorphisms affect the expression levels of various genes and thus affect various traits of the individual. Different alleles can thus have different effects on phenotypes. If the phenotype caused by a particular SNP allele improves an individual's chances of survival and reproduction (in comparison to other individuals within a population), then DNA inheritance mechanisms imply that the responsible allele, and therefore the phenotype, will be preferentially passed on to the next generation. This process, by which inheritable traits that confer a differential reproductive advantage become more common in populations over generations, is known as natural selection. Even though natural selection directly acts on traits or phenotypes, it indirectly affects allele frequencies.

Recombination is the process by which genomic segments from the two haplotypes of an individuals chromosomes can be rearranged to form gametes with new genetic sequences. When recombination occurs between two loci, it often decouples the alleles present at these loci to create new patterns of allelic coupling. This leads to varying degrees of coupling between the alleles at loci that are physically close to each other on the chromosome. This pattern is called linkage disequilibrium and is a common way of studying genetic variation [Durbin et al., 2010].

Various demographic processes such as population growth and contractions affect genetic diversity at the population. Migration is another important process that affects genetic variation in populations. When genetically different populations encounter each other through migration and produce offspring, the chromosomes in the resulting population contain genetic contributions from both of the ancestral populations. This process of genetic mixing is called admixture and the resulting population is referred to as an admixed population.

2.2 Population structure

Population structure is the presence of genetic similiarities and differences within and between groups of individuals. This is a problem of long-standing interest for reconstructing the ancestral history of modern populations using DNA polymorphisms [Cavalli-Sforza et al., 1994]. Genetic population structure can shed light on the evolutionary history and migrations of modern populations [Bowcock et al., 1994, Conrad et al., 2006, Rosenberg et al., 2002]. It also provides guidelines for more accurate association studies [Roeder et al., 1998] and is useful for many other population genetics problems [Hammer et al., 1998, Queller et al., 1993, Templeton, 2002].

Early attempts at recovering population structure used distance-based phylogenetic methods on genotype data [Bowcock et al., 1994]. While these methods easy to apply and interpret visually, they have important disadvantages: the clustering obtained can be heavily dependent on the distance measure used; and it is difficult to estimate the confidence of the resulting clustering. These methods were therefore replaced by model-based clustering approaches which modeled genomes as a mixture of contributions from ancestral populations. The earliest method in this class of approaches is the *Structure* model by Pritchard et al. [2000b]. A number of other model-based methods have been proposed to address the population structure problem in various contexts, such as the NewHybrids program [Anderson and Thompson, 2002] for classifying species hybrids into categories and the BAPS program [Corander et al., 2003] to find the best partition of a set of individuals into sub-populations on the bases of genotypes. Another class of methods to detect population structure has been developed using the eigenanalysis framework [Patterson et al., 2006]. *Eigensoft* uses eigen-decomposition methods to project individual genotypes into low-dimensional subspaces that can be used to visualize and examine genetic population structure. These methods provide formal tests for statistical significance of the population structure and are very efficient and scalable.

In this thesis, we develop model-based methods that take into account the evolutionary processes that shape genetic variation in populations. We therefore examine in more detail the

Structure method which forms the basis of these extensions.

2.3 The *Structure* model

The *Structure* model by Pritchard et al. [2000b] provides a way of probabilistically clustering individuals into groups called ancestral populations. It is a model-based approach which uses a statistical methodology known as the allele-frequency admixture model to stratify population structures. This model, and admixture models in general arising in genetic and other contexts (for instance, in document modeling [Blei et al., 2003]), belong to a more general class of hierarchical Bayesian models known as the *mixed membership models* [Erosheva et al., 2004]. Such a model postulates that the ensemble of genetic markers of an individual, is made up of independently and identically distributed (*iid*) instantiations [Pritchard et al., 2000b] from multiple population-specific fixed-dimensional multinomial distributions (known as *allele frequency profiles* [Falush et al., 2003], or AP) of marker alleles. Under this assumption, the admixture model identifies each ancestral population by a specific AP (that defines a unique vector of allele frequencies of each marker in each ancestral population), and displays the fraction of contributions from each AP in a modern individual genome as an *admixing vector* (also known as an *ancestral proportion vector* or *structure vector*) in a *structural map* over the population sample in question. Figure 2.1 shows an example of a structural map of four modern populations inferred from a portion of the HapMap multi-population dataset by *Structure*. In this *population structural map*, the *admixing vector* underlying each individual is represented as a thin vertical line of unit length and multiple colors, with the height of each color reflecting the fraction of the individual's genome originated from a certain ancestral population denoted by that color and formally represented by a unique AP. This method has been applied to the HGDP-CEPH Human Genome Diversity Cell Line Panel in Rosenberg et al. [2002] and to many other studies, and has unraveled interesting patterns in the genetic structures of world population.

In this section, we examine the generative process underlying the *Structure* model and how

it represents ancestral populations for ancestry inference.

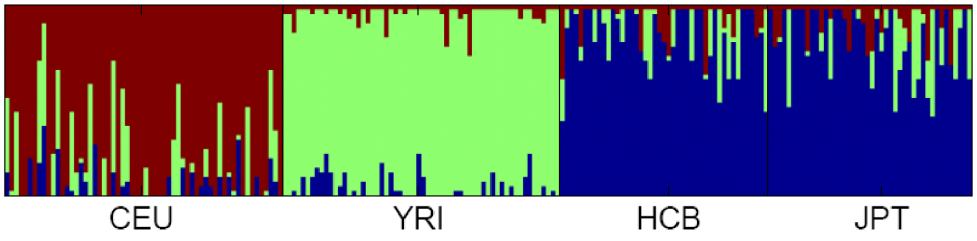


Figure 2.1: Population structural map inferred by *Structure* on HapMap data consisting of 4 populations.

2.3.1 Representation: Population-Specific *Allele Frequency* Profiles

Since all markers that are used for population strucure stratification are polymorphic in nature, it is not surprising that the most intuitive representation of an ancestral population is a set of frequency vectors for all alleles observed at all the loci. Specifically, we can represent an ancestral population k by a unique set of population-specific *multinomial* distributions $\boldsymbol{\beta}^k \equiv \{\vec{\beta}_i^k \; ; \; i = 1 : I\}$, where $\vec{\beta}_i^k = [\beta_{i,1}^k, \ldots, \beta_{i,L_i'}^k]$ is the vector of multinomial parameters, also known as an *allele frequency profile* [Falush et al., 2003], or AP, of the allele distribution at locus i in ancestral population k; L_i' denotes the total number of observed marker alleles at locus i, and I denotes the total number of marker loci. This representation, known as *population-specific allele frequency profiles*, is used by the program *Structure*.

2.3.2 Generative process

For example, for every individual, the alleles at all loci may be inherited from founders in different ancestral populations, each represented by a unique distribution of founding alleles and the way they can be inherited. Formally, this scenario can be captured in the following generative process:

1. For each individual n, draw the admixing vector: $\vec{\theta}_n \sim P(\cdot|\alpha)$, where $P(\cdot|\alpha)$ is a prechosen structure prior.

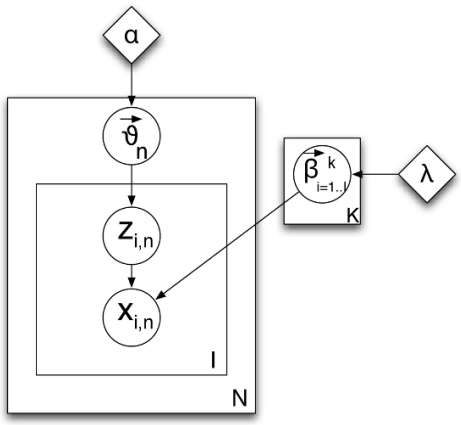


Figure 2.2: Graphical model for *Structure*: the circles represent random variables and diamonds represent hyperparameters.

2. For each marker allele $x_{i,n_e} \in \mathbf{x}_n$

   - 2.1: draw the latent *ancestral-population-origin* indicator $z_{i,n_e} \sim \text{Multinomial}(\cdot|\vec{\theta}_n)$
   - 2.2: draw the allele $x_{i,n_e}|z_{i,n_e} = k \sim P_k(\cdot|\Theta_i^k)$.

In *Structure*, the ancestral populations are represented by a set of population-specific allele frequency profiles (APs). Thus the distribution $P_k(\cdot|\Theta^k)$ from which an observed allele can be sampled is a multinomial distribution defined by the frequencies of all observed alleles in the ancestral population, i.e., $x_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}_i^k)$. Using this probability distribution in the general admixture scheme outlined above, we can see that *Structure* essentially implements an *admixture of population-specific allele frequency profiles* model. Figure 2.2 shows the graphical model representing the *Structure* generative process. This model has been successfully applied to human genetic data in Rosenberg et al. [2002], and it has been generalized to allow linked loci and correlated allele frequencies in Falush et al. [2003].

2.4 Disease association

Disease association is the task of inferring genetic variants that contribute to disease risk or explain phenotypic diversity in inheritable traits. Traditional methods for genetic analysis of diseases used techniques such as linkage analysis of candidate markers or genes and quantitative trait locus (QTL) mapping using one marker and one phenotype at a time [Easton et al., 1993]. Recent methods allow analysis of multiple markers simultaneously [Balding, 2006]. Methods such as eigenanalysis [Price et al., 2006] and regression [Cordell and Clayton, 2002] can perform simultaneous analysis of multiple markers for association. Mixed models such as EMMA [Kang et al., 2008] extend the regression framework to model the association problem (with confounding variables) as a linear mixed model.

Genetic association studies are usually set up in one of three different ways:

Familial studies In familial studies, pedigrees with a known history of a particular disease are genotyped. This avoids the problem of population stratification (an important problem occuring in association studies that will be explained in more detail in Section 2.4.1). However, the restriction on the individuals that can be included in the study limits the power of the method in finding associations.

Case-control design Case-control studies involve a comparison between the genotypes of two sets of individuals characterized by presence or absence of the phenotype of interest. Cases are a group of individuals who exhibit the phenotype of interest (a diease or a complex trait). Controls are individuals who do not show prevalence of the phenotype. The underlying assumption is that genotypic differences (in terms of the frequency of certain allelic variants) between cases and controls are likely to be at markers which are causally related to the phenotype being studied. In most recent studies, association studies are set up using a case-control design.

Population cohorts Rather than designate two different sets of individuals as cases and controls, population cohorts follow a single set of individuals over a longer period of time,

collecting phenotypic information for multiple traits. This limits the number of "cases" for a particular disease that might be present in the cohort, but the resulting data includes a lot of longitudinal information about multiple phenotypes that can be useful for other studies of diseases. It can be used to study the effect of epigenetic factors [Wong et al., 2004] and pleiotropy [Cordell and Clayton, 2005].

For diseases such as age-related macular degeneration, it has been found that only a few common variants having large effects account for most of the heritability of the trait. Scenarios such as these are conducive to analysis by genome-wide association studies. In many other diseases, most common variants only add small increments to the disease risk and explain only a small percentage of heritability. An example of such a trait is human height, with an estimated heritability of 80%. Genome-wide association studies have indicated $\sim$40 loci that might be associated with human height, but they explain only 5% of the phenotypic variance of human height. Similar problems have been encountered when trying to explain the heritability of other complex traits using association studies. Below we discuss some more of the challenges that are faced when performing association studies.

2.4.1 Challenges in genetic association studies

Population stratification Case-control studies are based on the assumption that genotype differences between cases and controls are likely to be causually related to the phenotype. However, if there is unidentified population stratification between the cases and controls, this assumption does not hold true. If the cases disproportionately represent a genetic population in comparison to the controls, then any SNP with allele frequencies differing between the cases and controls will (incorrectly) be found to be associated with the phenotype, when it is only truly associated with distinguishing case or control status. A variety of methods have been proposed to identify and correct for population stratification in association studies [Price et al., 2006, Pritchard et al., 2000b, Puniyani et al., 2010, Roeder et al., 1998].

Insufficient sample size It has been suggested that the partial success of genetic association studies could be a result of not sampling enough individuals. Small sample sizes could result in rigorous tests of statistical significance failing to identify variants of small or moderate effects as causal. Recent work by Yang et al. [2010] suggests that increasing sample sizes identifies new SNPs that allow us to explain up to 40% of the heritability of human height. While this is a significant improvement, it still accounts for only half of the estimated heritability of the trait.

Single locus association statistics Many traditional tests for association are single-locus tests for statistical significance. Due to the large number of statistical tests that have to be performed for all genotyped SNPs, a correction factor must be applied to the test statistic to avoid false positives. A commonly used correction is the Bonferroni correction, by which the test statistic is reduced by a factor of the number of SNPs. This assumes that all the tests performed are independent. However, due to linkage disequilibrium, SNPs are correlated and therefore the tests are not independent of each other. The Bonferroni correction, therefore, is too conservative and ignores weak associations.

Effect size distribution The early genome-wide association studies have been able to identify candidate SNPs that have large effects. The undiscovered causual variants are likely to have smaller effects. Therefore finding newer candidate loci in association studies is likely to be a harder problem [Park et al., 2010].

Comon disease, rare variants Current SNP chips capture variation only at loci where the minor allele frequency (MAF) is between 1-5%. However, low frequency (MAF $\leq$ 1%) variants and rare variants (MAF $\leq$ 0.01 %) are not captured. Since many traits are multifactorial, a relatively small number of rare variants with moderate effect could account for a large percentage of the trait heritability.

Chapter 3

Population structure in the presence of admixture and allele mutations

Traditional methods for analyzing population structure, such as the *Structure* program [Pritchard et al., 2000b], ignore the influence of the effect of allele mutations between the ancestral and current alleles of genetic markers, which can dramatically influence the accuracy of the structural estimation of current populations. A study of these effects can also reveal additional information about population evolution such as the the divergence time and migration history of admixed populations. We propose *mStruct* [Shringarpure and Xing, 2009], an admixture of population-specific mixtures of inheritance models , to address the task of structure inference and mutation estimation jointly through a hierarchical Bayesian framework. We develop a variational algorithm for performing inference for the model. We validate our method on simulated data, and use it to analyze the HGDP-CEPH cell line panel of microsatellites used in Rosenberg et al. [Rosenberg et al., 2002] and the HGDP SNP data used in Conrad et al. [Conrad et al., 2006]. We present a comparison of the structural maps of world populations estimated by *mStruct* and *Structure* and report potentially interesting mutation patterns in world populations estimated by *mStruct*.

3.1 Introduction

The recent deluge of genomic polymorphism data has fueled the long-standing interest in the analysis of patterns of genetic variations to reconstruct the ancestral structures of modern human populations. Various methods have been proposed for detecting population structures based on multi-locus genotype information from a set of indviduals, starting with the *Structure* model by Pritchard et al. [2000b]. However, even though *Structure* was originally built on a genetic admixture model, in reality the structural patterns derived by *Structure* in various studies often turn out to be distinct clusters amongst the study populations (e.g., Figure 2.1), which has led many to think of it as a clustering program rather than a tool for uncovering genetic admixing as was originally intended. This issue motivated us to develop a new approach to analyze admixed genetic samples.

An extension of *Structure*, known as Structurama [Huelsenbeck and Andolfatto, 2007, Pella and Masuda, 2006], relaxes the finite dimensional assumption on ancestral populations in the admixture model by employing a Dirichlet process prior over the ancestral allele frequency profiles. This allows automatic estimation of the maximum *a posteriori* probable number of ancestral populations. This extension is a useful improvement since it eliminates the need for manual selection of the number of ancestral populations. Anderson and Thompson [2002] address the problem of classifying species hybrids into categories using a model-based Bayesian clustering approach implemented in the NewHybrids program. While this problem is not exactly identical to the problem of stratifying the structure of highly admixed populations, it is useful for structural analysis of populations that were recently admixed. The BAPS program developed by Corander et al. [2003] also uses a Bayesian approach to find the best partition of a set of individuals into sub-populations on the basis of genotypes. Parallel to the aforementioned model-based approaches for genomic structural analysis, direct algebraic eigen-decomposition and dimensionality reduction methods, such as the *Eigensoft* program developed by Patterson et al. [2006] based on Principal Components Analysis (PCA), offer an alternative approach to explore and

visualize the ancestral composition of modern populations, and facilitate formal statistical tests for significance of population differentiation. However, unlike the model-based methods such as the *Structure*, where each inferred ancestral population bears a concrete genetic meaning as a population-specific allele-frequency profile, the eigen-vectors computed by *Eigensoft* represent the mutually-orthogonal directions in an abstract low-dimensional ancestral space, in which population samples can be embedded and visualized. These eigen-vectors can be understood as mathematical surrogates of independent genetic sources underlying a population sample, but lack a concrete interpretation under a generative genetic inheritance model (from here on, we will use the term "inheritance model" to describe the process by which a descendant allele is derived from an ancestral allele). Analyses based on *Eigensoft* are usually limited to 2-dimensional ancestral spaces, offering limited power in stratifying highly admixed populations.

This progress notwithstanding, an important aspect of population admixing that is largely missing in the existing methods is the effect of allele mutations between the ancestral and current alleles of genetic markers, which can dramatically influence the accuracy of the structural estimation of current populations. It can also reveal additional information about population evolution, such as the relative divergence time and migration history of admixed populations.

Consider for example the *Structure* model. Since an AP merely represents the *frequency* of alleles in an ancestral population rather than the actual allelic content or haplotypes of the alleles themselves, the admixture models developed so far based on AP do not model genetic changes due to mutations from the ancestral alleles. Indeed, a serious pitfall of the model underlying *Structure*, as pointed out in Excoffier and Hamilton [2003], is that there is no mutation model for modern individual alleles with respect to hypothetical common prototypes in the ancestral populations. That means, every unique allele in the modern population is assumed to have a distinct ancestral proportion, rather than allowing the possibility of it just being a descendant of some common ancestral allele that can also give rise to other closely related alleles at the same locus of other individuals in the modern population. Thus, while *Structure* aims to provide ancestry

information for each individual and each locus, there is no explicit representation of the "ancestors" as a physical set of "founding alleles". Therefore, the inferred population structural map emphasizes revealing the contributions of *abstract* population-specific ancestral proportion profiles, which does not necessarily reflect individual diversity or the extent of genetic changes with respect to the founders. Due to this limitation, *Structure* does not enable inference of the founding genetic patterns, the age of the founding alleles, or the population divergence time [Excoffier and Hamilton, 2003].

The lack of an appropriate allele mutation model in a structural inference program can also compromise our ability to reliably assess the amount or level of genetic admixing in different populations. The *Structure* model, like several other related models [Blei et al., 2003], is based on the fundamental assumption of the presence of genetic admixing among multiple founding populations. However, as we shall see later, on real population data such as the HGDP-CEPH panel, it produces results that favor clustering individuals into predominantly one allele frequency profile or another, thus leading us to conclude that there was little or no admixing between the ancestral human populations. We believe that this occurs due to the absence of a mutation model in *Structure*. While a partitioning of individuals would be desirable for clustering them into groups, it does not offer enough biological insight into the intermixing of the populations.

We develop *mStruct* (which stands for *Structure under mutations*), based on a new model: an admixture of population-specific mixtures of inheritance models (AdMim). Statistically, AdMim is an *admixture of mixture models*, which represents each ancestral population as a mixture of ancestral alleles each with its own inheritance process, and each modern individual as an "ancestry vector" (or *structure vector*) that reflects membership proportions of the ancestral populations. As we explain shortly, *mStruct* facilitates estimation of both the *structural map* of populations (incorporating mutations) and the mutation parameters of either SNP or microsatellite alleles under various contexts. We develop a new variational inference algorithm for estimating the structure vectors and other model parameters of interest. We compare our method with *Structure*

on simulated genotype data, and on the microsatellite and SNP genotype data of world populations [Conrad et al., 2006, Rosenberg et al., 2002]. Our results using microsatellite data reveal the presence of significant levels of genetic admixing among the founding populations underlying the HGDP-CEPH cell line panel, as well as consequences of expansion of humans out of Africa. Our results suggest that the inability of *Structure* to model mutations during genetic admixing could have caused it to detect correct clustering but very low levels of genetic admixing in each modern population in the HGDP-CEPH data. We also report interesting visualizations of genetic divergence in world populations revealed by the mutation patterns estimated by *mStruct*.

3.2 The statistical model

Although both *mStruct* and *Structure* are mixed-membership models, the *mStruct* model differs from the *Structure* model in two main aspects: the representation of ancestral populations, and the generative process for sampling a modern individual from the ancestral populations. In this section we describe in detail the statistical underpinning of these two aspects.

3.2.1 Representation: Population-Specific *Mixtures of Ancestral Alleles*

An AP does not enable us to model the possibility of mutations, i.e., there is no way of representing a situation where two observed alleles might have been derived from a single ancestral allele by two different mutations. This possibility can be represented by a genetically more realistic statistical model known as the *population-specific mixture of ancestral alleles* (MAA). For each locus i, an MAA for ancestral population k is a set $\Theta_i^k \equiv \{\mu_i^k, \delta_i^k, \vec{\beta}_i^k\}$ consisting of three components: 1) a set of *ancestral* (or founder) alleles $\mu_i^k \equiv \{\mu_{i,1}^k, \ldots, \mu_{i,L_i}^k\}$, which can differ from their descendent alleles in the modern population; 2) a mutation parameter δ_i^k associated with the locus, which can be further generalized to be allele-specific if necessary; and 3) an AP $\vec{\beta}_i^k$ which now represents the frequencies of the *ancestral* alleles. Here L_i denotes the

total number of *ancestral* alleles at loci i, which is different from L_i' in the previous subsection, which denotes the total number of *observed* alleles at loci i. By explictly associating a mutation model with an ancestral population, we can now capture mutation events as described above. It is important to note that the mutation parameter δ is not the mutation rate commonly referred to in literature. As we shall see later, it is a measure of the variability of a locus which can be described approximately as the combined effect of the per-generation mutation rate and the age of the population.

An MAA is strictly more expressive than an AP, because the incorporation of a mutation model helps to capture details about the population structure which an AP cannot; and the MAA reduces to the AP when the mutation rates (and hence the mutation parameters) become zero and the founders are identical to their descendents. MAA is also arguably more realistic because it allows mutation rates (and mutation parameters) to be different for different founder alleles even within the same ancestral population, as is commonly the case with many genetic markers. For example, the mutation rates for microsatellite alleles are believed to be dependent on their length (number of repeats). As we shall show shortly, with an MAA, one can examine the mutation parameters corresponding to each ancestral population via Bayesian inference from genotype data; this might enable us to infer the age of alleles, and also estimate population divergence times subject to a calibration constant.

Let $i \in \{1, \ldots, I\}$ index the position of a locus in the study genome, $n \in \{1, \ldots, N\}$ index an individual in the study population, and $e \in \{0, 1\}$ index the two possible parental origins of an allele (in this study we do not require strict phase information of the two alleles, so the index e is merely used to indicate ploidy of the data). Under an MAA specific to an ancestral population k, the correspondence between a marker allele X_{i,n_e} and a founder $\mu_{i,l}^k \in \mu_i^k$ is not directly observable. For each allele founder $\mu_{i,l}^k$, we associate with it an inheritance model $p(\cdot|\mu_{i,l}^k, \delta_{i,l}^k)$ from which descendants can be sampled. Then, given specifications of the ancestral population from which X_{i,n_e} is derived, which is denoted by hidden indicator variable Z_{i,n_e}, the conditional

distribution of X_{i,n_e} under MAA follows a mixture of population-specific inheritance models:

$$P(x_{i,n_e} = l' \mid z_{i,n_e} = k) = \sum_{l=1}^{L} \beta_{i,l}^k P(x_{i,n_e} | \mu_{i,l}^k, \delta_{i,l}^k). \tag{3.1}$$

Comparing to the counterpart of this function under AP: $P(x_{i,n_e} = l' \mid z_{i,n_e} = k) = \beta_{i,l'}^k$, we can see that the latter cannot explicitly model allele diversities in terms of molecular evolution from the founders.

3.2.2 Generative process

We propose to represent each ancestral population by a set of population-specific MAAs. Recall that in an MAA for each locus we define a finite set of founders with prototypical alleles $\mu_i^k \equiv \{\mu_{i,1}^k, \ldots, \mu_{i,L_i}^k\}$ that can be different from the alleles observed in a modern population; each founder is associated with a unique frequency $\beta_{i,l}^k$, and a unique (if desired) mutation model from the prototype allele parameterized by rate $\delta_{i,l}^k$. Under this representation, now the distribution $P_k(\cdot|\Theta_i^k)$ from which an observed allele can be sampled becomes a mixture of inheritance models each defined on a specific founder; and the ensuing sampling module that can be plugged into the general admixture scheme outlined in Section 2.3.2 becomes a two-step generative process. The entire generative process can be written as:

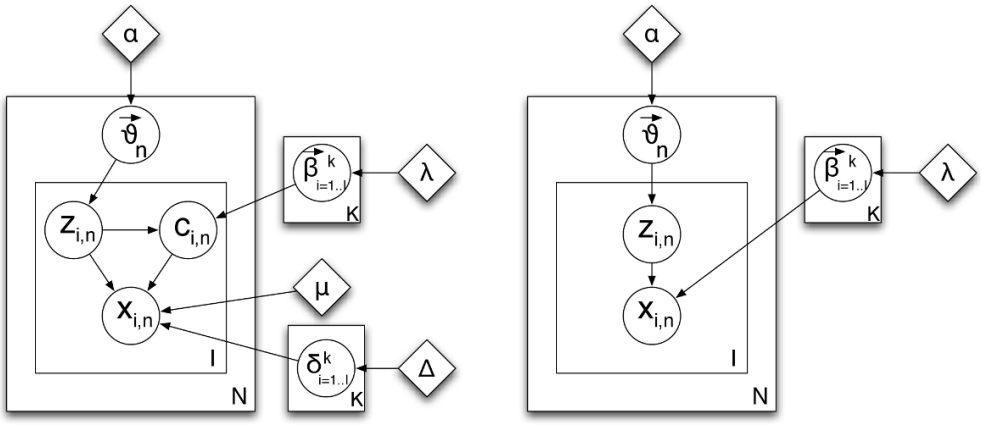
1. For each individual n, draw the admixing vector: $\vec{\theta}_n \sim P(\cdot|\alpha)$, where $P(\cdot|\alpha)$ is a pre-chosen structure prior.

2. For each marker allele $x_{i,n_e} \in \mathbf{x}_n$

   - 2.1: draw the latent *ancestral-population-origin* indicator $z_{i,n_e} \sim \text{Multinomial}(\cdot|\vec{\theta}_n)$
   - 2.2a: draw the latent founder indicator $c_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}_i^k)$;
   - 2.2b: draw the allele $x_{i,n_e}|c_{i,n_e} = l, z_{i,n_e} = k \sim P_m(\cdot|\mu_{i,l}^k, \delta_{i,l}^k)$,

where $P_m()$ is a mutation model that can be flexibly defined based on whether the genetic markers are microsatellites or single nucleotide polymorphisms. We call this model an *admixture of*

population-specific inheritance models (AdMim), whereas the *Structure* model is only an *admixture of population-specific allele frequency profiles*. Figure 3.1(a) shows a graphical model representation of the overall generative scheme for AdMim, in comparison with the admixture of population-specific allele rates discussed earlier. From the figure, we can clearly see that *mStruct* is an extended *Structure* model which allows copying errors.



(a) *mStruct* (b) *Structure 2.1*

Figure 3.1: Graphical Models: the circles represent random variables and diamonds represent hyperparameters.

For simplicity of presentation, in the model described above, we assume that for a particular individual, the genetic markers at each locus are conditionally *iid* samples from a set of population-specific fixed-dimensional mixture of inheritance models, and that the set of founder alleles (but not their frequencies) at a particular locus is the same for all ancestral populations (i.e., $\mu_i^k \equiv \mu_i$). We shall also assume that the mutation parameters for each population at any locus are independent of the alleles at that locus (i.e., $\delta_{i,l}^k \equiv \delta_i^k$). Also, our model assumes Hardy-Weinberg equilibrium within populations. The simplifying assumptions of *unlinked loci* and *no linkage disequilibrium between loci within populations* can be easily removed by incorporating Markovian dependencies over ancestral indicators Z_{i,n_e} and Z_{i+1,n_e} of adjacent loci, and over other parameters such as the allele frequencies $\vec{\beta}_i^k$ in exactly the same way as in *Structure*. We can also introduce Markovian dependencies over mutation parameters at adjacent loci, which

might be desirable to better reflect the dynamics of molecular evolution in the genome. We defer such extensions to future work.

3.2.3 Mutation model

As described above, our model is applicable to data for almost all kinds of genetic markers by plugging in an appropriate allele mutation model (i.e., inheritance model) $P_m()$. We now discuss mutation models for microsatellites and SNPs.

Microsatellite mutation model

Microsatellites are a class of tandem-repeat loci that involve a DNA unit that is $1-4$ basepair in length. Microsatellite DNA has significantly high mutation rates as compared to other DNA, with mutation rates as high as 10^{-3} or 10^{-4} [Henderson and Petes, 1992, Kelly et al., 1991]. The large amount of variations present in microsatellite DNA make it ideal for differentiating founder patterns between closely related populations. Microsatellite loci have been used before DNA fingerprinting [Queller et al., 1993], linkage analysis [Dietrich et al., 1992], and in the reconstruction of human phylogeny [Bowcock et al., 1994]. By applying theoretical models of microsatellite evolution to data, questions such as time of divergence of two populations can be attempted to be addressed [Pisani et al., 2004, Zhivotovsky et al., 2004].

The choice of a suitable microsatellite mutation model is important, for both computation and interpretation purposes. Below we discuss the mutation model that we use and the biological interpretation of the parameters of the mutation model. We begin with a stepwise mutation model for microsatellites widely used in forensic analysis [Lin et al., 2006, Valdes et al., 1993].

This model defines a conditional distribution of a progeny allele b given its progenitor allele a, both of which take continuous values:

$$p(b|a) = \frac{1}{2}\xi(1-\delta)\delta^{|b-a|-1}, \tag{3.2}$$

where ξ is the mutation rate (probability of any mutation), and δ is the factor by which mutation decreases as distance between the two alleles increases. Although this mutation distribution is not stationary (i.e., it does not ensure allele frequencies to be constant over the generations), it is commonly used in forensic inference due to its simplicity. To some degree δ can be regarded as a parameter that controls the probability of unit-distance mutation, as can be seen from the following identity: $p(b+1|a)/p(b|a) = \delta$.

In practice, the alleles for almost all microsatellites are represented by discrete counts. The two-parameter stepwise mutation model described above complicates the inference procedure. We propose a discrete microsatellite mutation model that is a simplification of Eq. 3.2, but captures its main idea. We posit that: $P(b|a) \propto \delta^{|b-a|}$. Since $b \in [1, \infty)$, the normalization constant of this distribution is:

$$\begin{aligned} \sum_{b=1}^{\infty} P(b|a) &= \sum_{b=1}^{a} \delta^{a-b} + \sum_{b=a+1}^{\infty} \delta^{b-a} \\ &= \frac{1-\delta^a}{1-\delta} + \frac{\delta}{1-\delta} \\ &= \frac{1+\delta-\delta^a}{1-\delta}, \end{aligned}$$

which gives the mutation model as

$$P(b|a) = \frac{1-\delta}{1-\delta^a+\delta}\delta^{|b-a|}. \tag{3.3}$$

We can interpret δ as a variance parameter, the factor by which probability drops as a fuction of the distance between the mutated version b of the allele a. Figure 3.2 shows the discrete pdf for various values of δ.

SNP mutation model

SNPs, or single nucleotide polymorphisms, represent the largest class of individual differences in DNA. In general, there is a well-defined correlation between the age of the mutation producing a SNP allele and the frequency of the allele. For SNPs, we use a simple pointwise mutation model, rather than more complex block models. Thus, the observations in SNP data are only

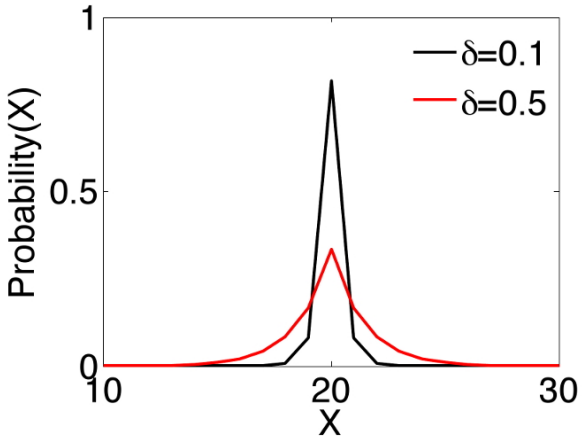


Figure 3.2: Discrete pdf for two values of mutation parameter.

binary (0/1) in nature. Thus, given the observed allele b, we say that the probability of it being derived from the founder allele a is given by:

$$P(b|a) = \delta^{\mathcal{I}[b=a]} \times (1-\delta)^{\mathcal{I}[b\neq a]}; \quad a, b \in \{0, 1\}. \tag{3.4}$$

In this case, the mutation parameter δ is the probability that the observed allele is not identical to the founder allele, but derived from it due to a mutation.

Other modeling issues

In our model description, we defined an ancestral population using a set of founder alleles. To use the model to analyze data, we need to decide how these founder alleles can be obtained. Below, we describe how this can be accomplished. We also explain how to define a prior on the mutation parameter to enforce model identifiability.

Determination of founder set at each locus: According to our model assumptions, there can be a different number of founder alleles at each locus. This number is typically smaller than the number of alleles observed at each marker since the founder alleles are "ancestral". To estimate the appropriate number and allele states of founders, we fit finite mixtures (of fixed size, corresponding to the desired number of ancestral alleles) of microsatellite mutation models over

all the measurements at a particular marker for all individuals. We use the Bayesian Information Criterion (BIC) [Schwarz, 1978] to determine the best number and states of founder alleles to use at each locus, since information criteria tend to favor smaller number of founder alleles which fit the observed data well.

For each locus, we fit many different finite-sized mixtures of mutation distributions, with the size varying from 1 to the number of observed alleles at the locus. For each mixture size, the likelihood is optimized and a BIC value is computed. The number of founder alleles is chosen to be the size of the mixture that has the best (minimum) BIC value. We can do this as a pre-processing step before the actual inference or estimation procedures. This is possible since we assumed that the set of founder alleles at each locus was the same for all populations.

Choice of mutation prior: In our model, the δ parameter, as explained above, is a population-specific parameter that controls the probability of stepwise mutations. Being a parameter that controls the variance of the mutation distribution, there is a possibility that inference on the model will encourage higher values of δ to improve the log-likelihood, in the absence of any prior distribution on δ. To avoid this situation, and to allow more meaningful and realistic results to emerge from the inference process, we impose on δ a beta prior that will be biased towards smaller values of δ. The beta prior will be a fixed one and will not be among the parameters we estimate.

3.3 Inference

For notational convenience, we will ignore the diploid nature of observations in the analysis that follows. With the understanding that the analysis is carried out for an arbitrary n^{th} individual, we will drop the subscript n. Also, we overload the indicator variables z_i and c_i to be both, arrays with only one element equal to 1 and the rest equal to 0, as well as scalars with a value equal to the index at which the array forms have 1s. In other words: $z_i \in \{1, \ldots, K\}$ or $z_i =$

$[z_{i,1}, \ldots, z_{i,K}]$, where $z_{i,k} = \mathcal{I}[z_i = k]$, and $\mathcal{I}[\cdot]$ denotes an indicator function that equals to 1 when the predicate argument is true and 0 otherwise. A similar overloading is also assumed for the c_i variables. For generalization across different types of markers, we shall use $f(x_i|\mu_{i,c_i}, \delta_{i,z_i})$ to denote $P(x_i|c_i, z_i, \mu_i, \delta_i)$. Different mutation models can be used in AdMim by varying the form of the function $f()$.

The joint probability distribution of the the data and the relevant variables under the AdMim model can then be written as:

$$
\begin{aligned}
& P\left(\mathbf{x}, \mathbf{z}, \mathbf{c}, \vec{\theta} | \alpha, \boldsymbol{\beta}, \mu, \delta\right) \\
&= p\left(\vec{\theta}|\alpha\right) \prod_{i=1}^{I} P\left(z_i|\vec{\theta}\right) P\left(c_i|z_i, \vec{\beta}_i^{k=1:K}\right) P\left(x_i|c_i, z_i, \mu_i, \delta_i^{k=1:K}\right).
\end{aligned}
$$

The marginal likelihood of the data can be computed by summing/integrating out the latent variables:

$$
\begin{aligned}
P\left(x|\alpha, \boldsymbol{\beta}, \mu, \delta\right) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \int \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right) \ldots \\
\times \prod_{i=1}^{I} \sum_{k=1}^{K} \left(\prod_{k=1}^{K} \theta_k^{z_{i,k}}\right) \sum_{i=1}^{I} \prod_{k=1}^{K} \prod_{l=1}^{L_i} \left(\beta_{i,l}^{k}\right)^{c_{i,l} z_{i,k}} \ldots \\
\times P\left(x_i|\mu_{i,l}, \delta_i^k\right)^{c_{i,l} z_{i,k}} d\vec{\theta}.
\end{aligned}
$$

However, a closed-form solution to this summation/integration is not possible, and indeed exact inference on hidden variables such as the map vector $\vec{\theta}$, and estimation of model parameters such as the mutation rates δ under AdMim is intractable. Pritchard et al. [2000a] developed an MCMC algorithm for inference for their admixture model underlying *Structure*. While it is straightforward to implement a similar MCMC scheme for AdMim, we choose to apply an approximate inference method known as variational inference [Jordan et al., 1999] for computational efficiency.

3.3.1 Variational Inference

We use a mean-field approximation for performing inference on the model. This approximation method approximates an intractable joint posterior $p()$ of all the hidden variables in the model by a product of marginal distributions $q() = \prod q_i()$, each over only a single hidden variable. The optimal parameterization of $q_i()$ for each variable is obtained by minimizing the Kullback-Leibler divergence between the variational approximation q and the true joint posterior p. Using results from the Generalised Mean Field theory [Xing et al., 2003], we can write the variational distributions of the latent variables in AdMim as follows:

$$
\begin{aligned}
q(\vec{\theta}) &\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + \sum_{i=1}^{I} \langle z_{i,k} \rangle} \\
q(c_i) &\propto \prod_{l=1}^{L} \left(\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle} \right)^{c_{i,l}} \\
q(z_i) &\propto \prod_{k=1}^{K} \left(e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right) \right)^{z_{i,k}}.
\end{aligned}
$$

In the distributions above, the '$\langle \cdot \rangle$' are used to indicate the expected values of the enclosed random variables. A close inspection of the above formulas reveals that these variational distributions have the form $q(\vec{\theta}) \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_K)$, $q(z_i) \sim \text{Multinomial}(\rho_{i,1}, \ldots, \rho_{i,K})$, and $q(c_i) \sim \text{Multinomial}(\xi_{i,1}, \ldots, \xi_{i,L})$, respectively, of which the parameters $\gamma_k, \rho_{i,k}$ and $\xi_{i,l}$ are given by the following equations:

$$
\begin{aligned}
\gamma_k &= \alpha_k + \sum_{i=1}^{I} \langle z_{i,k} \rangle \\
\rho_{i,k} &= \frac{e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right)}{\sum_{k=1}^{K} \left(e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right) \right)} \\
\xi_{i,l} &= \frac{\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle}}{\sum_{k=1}^{K} \left(\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle} \right)}
\end{aligned}
$$

and they have the properties: $\langle \log(\theta_k) \rangle = \psi(\gamma_k) - \psi(\sum_k \gamma_k)$, $\langle z_{i,k} \rangle = \rho_{i,k}$ and $\langle c_{i,l} \rangle = \xi_{i,l}$,

which suggest that they can be computed via fixed point iterations. (The digamma function $\psi()$ used above is the first derivative of the logarithm of the gamma function $\Gamma()$.) It can be shown that this iteration will converge to a local optimum, similar to what happens in an EM algorithm. Empirically, a near global optimal can be obtained by multiple random restarts of the fixed point iteration. Typically, such a mean-field variational inference converges much faster than sampling [Xing et al., 2003] (though we note that the sampling yields the full posterior while variational inference produces a unimodal approximation to the posterior). Upon convergence, we can easily compute an estimate of the map vector $\vec{\theta}$ for each individual from $q(\vec{\theta})$.

3.4 Parameter Estimation

The parameters of our model are the centroids μ, the mutation parameters δ, the ancestral allele frequency distributions $\boldsymbol{\beta}$, and the Dirichlet hyperparameter that is the prior on ancestral populations, α. For the hyperparameter estimation, we perform empirical Bayes estimation using the variational Expectation Maximization algorithm described in [Blei et al., 2003]. The variational inference described in Section 3.3.1 provides us with a tractable lower bound on the log-likelihood as a function of the current values of the hyperparameters. We can thus maximize it with respect to the hyperparameters. If we alternately carry out variational inference with fixed hyperparameters, followed by a maximization of the lower bound with respect to the hyperparameters for fixed values of the variational parameters, we can get an empirical Bayes estimate of the hyperparameters. The derivation leads to the following iterative algorithm:

1. (*E-step*) For each individual, find the optimizing values of the variational parameters $(\gamma_n, \rho_n, \xi_n; n \in 1, \ldots, N)$ using the variational updates described above.
2. (*M-step*) Maximize the resulting variational lower bound on the likelihood with respect to the model parameters, namely $\alpha, \beta, \mu, \delta$.

The two steps are repeated until the lower bound on the log-likelihood converges. The details of estimation of each hyperparameter are explained below.

3.4.1 Variational lower bound on log-likelihood

Denote the original set of hyperparameters by

$$\mathbb{H} = \{\alpha, \boldsymbol{\beta}, \mu, \delta\} \tag{3.5}$$

and the variational parameters for the n^{th} individual by

$$\mathbb{V}_n = \{\gamma_n, \rho_n, \xi_n\} \tag{3.6}$$

The variational lower bound to the log-likelihood for the n^{th} individual is given by:

$$\begin{aligned} L_n(\mathbb{H}, \mathbb{V}_n) = \mathbb{E}_q[\log p(x_n, \vec{\theta}_n, z_{.,n}, c_{.,n}; \mathbb{H})] \\ - \mathbb{E}_q[\log q(\vec{\theta}_n, z_{.,n}, c_{.,n}; \mathbb{H}, \mathbb{V}_n)] \end{aligned} \tag{3.7}$$

The subscripts indicate the n^{th} individual. In the analysis below, we use $z_{.,n}$ to denote $\{z_{1,n}, \ldots, z_{I,n}\}$ and $c_{.,n}$ to represent $\{c_{1,n}, \ldots, c_{I,n}\}$. As described earlier, we partition the variational approximation as:

$$q(\vec{\theta}_n, z_{.,n}, c_{.,n}; \mathbb{H}, \mathbb{V}) = q(\vec{\theta}_n) \prod_{i=1}^{I} q(z_{i,n}) q(c_{i,n}) \tag{3.8}$$

So we can expand Equation 3.7 as

$$\begin{aligned} L_n(\mathbb{H}, \mathbb{V}_i) = \mathbb{E}_q[\log p(\vec{\theta}_n; \alpha)] + \mathbb{E}_q[\log p(z_{.,n}|\vec{\theta}_n)] + \mathbb{E}_q[\log p(c_{.,n}|z_{.,n})] \\ + \mathbb{E}_q[\log p(x_n|c_{.,n}, z_{.,n}, \boldsymbol{\beta})] - \mathbb{E}_q[\log q(\vec{\theta}_n)] - \mathbb{E}_q[\log q(z., n)] - \mathbb{E}_q[\log q(c., n)] \end{aligned} \tag{3.9}$$

The lower bound to the total data log-likelihood is

$$L(\mathbb{H}, \mathbb{V}) = \sum_{n=1}^{N} L_n(\mathbb{H}, \mathbb{V}_n)$$

which, on substituting from Equation 3.9 becomes

$$
\begin{aligned}
L(\mathbb{H}, \mathbb{V}) = & \sum_{n=1}^{N} \mathbb{E}_q[\log p(\vec{\theta}_n; \alpha)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(z_{.,n}|\vec{\theta}_n)] \\
& + \sum_{n=1}^{N} \mathbb{E}_q[\log p(c_{.,n}|z_{.,n})] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(x_n|c_{.,n}, z_{.,n}, \boldsymbol{\beta})] \\
& - \sum_{n=1}^{N} \mathbb{E}_q[\log q(\vec{\theta}_n)] - \sum_{n=1}^{N} \mathbb{E}_q[\log q(z_{.,n})] \\
& - \sum_{n=1}^{N} \mathbb{E}_q[\log q(c_{.,n})]
\end{aligned}
\tag{3.10}
$$

To compute $\mathbb{E}_q[\log p(\vec{\theta}_n; \alpha)]$ and $\mathbb{E}_q[\log q(\vec{\theta}_n)]$, we will use the properties of a Dirichlet distribution, which is an exponential family distribution. If $\theta \sim \text{Dir}(\alpha)$, then the exponential family representation of $p(\theta; \alpha)$ is given by:

$$
p(\theta; \alpha) = \exp\left[\left(\sum_{k=1}^{K} (\alpha_k - 1) \log \theta_k\right) + \log \Gamma\left(\sum_{k=1}^{K} \alpha_k\right) - \sum_{k=1}^{K} \log \Gamma\left(\alpha_k\right)\right] \tag{3.11}
$$

So the natural parameter of the Dirichlet is $\eta_k = \alpha_k - 1$ and the sufficient statistic is $T(\theta_k) = \log \theta_k$. The log normalization factor is $\sum_{k=1}^{K} \log \Gamma\left(\alpha_k\right) - \log \Gamma\left(\sum_{k=1}^{K} \alpha_k\right)$. For an exponential distribution, the derivative of the log normalization factor with respect to the natural parameter is equal to the expected value of the sufficient statistic. Using this fact, we get:

$$
\text{E}[\log \theta_k; \alpha] = \psi\left(\alpha_k\right) - \psi\left(\sum_{k} \alpha_k\right) \tag{3.12}
$$

where ψ is the digamma function, the first derivative of the log Gamma function. The remaining expectation terms in Equation 3.10 are expectations of multinomial parameters, and hence are easy to calculate.

Simplifying each term in Equation 3.10, we get

$$
\begin{aligned}
L(\mathbb{H},\mathbb{V}) = {} & N \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - N\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\alpha_k - 1\right)\left[\psi\left(\gamma_{n,k}\right) - \psi\left(\sum_{k=1}^{K}\gamma_{n,k}\right)\right] \\
& + \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\rho_{n,i,k}\left[\psi\left(\gamma_{n,k}\right) - \psi\left(\sum_{k=1}^{K}\gamma_{n,k}\right)\right] \\
& + \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{L_i}\xi_{n,i,l}\rho_{n,i,k}\log\beta_{il}^{k} \\
& + \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{L_i}\xi_{n,i,l}\rho_{n,i,k}\left[\log\left(1-\delta_i^k\right) + \left|x_{i,n}-\mu_{i,l}\right|\log\delta_i^k - \log\left(1+\delta_i^k - \left(\delta_i^k\right)^{\mu_{i,l}}\right)\right] \\
& - \sum_{n=1}^{N}\left[\log\Gamma\left(\sum_{k=1}^{K}\gamma_{n,k}\right) - \sum_{k=1}^{K}\log\Gamma\left(\gamma_{n,k}\right) + \sum_{k=1}^{K}\left(\gamma_{n,k}-1\right)\left[\psi\left(\gamma_{n,k}\right) - \psi\left(\sum_{k=1}^{K}\gamma_{n,k}\right)\right]\right] \\
& - \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{l=1}^{L_i}\xi_{n,i,l}\log\xi_{n,i,l} \\
& - \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\rho_{n,i,k}\log\rho_{n,i,k}
\end{aligned}
\tag{3.13}
$$

Each line in Equation 3.13 corresponds to an expectation term in Equation 3.10. In the following subsections, we will briefly describe how the maximum-likelihood estimates of the hyperparameters were obtained from the variational lower bound.

3.4.2 Estimating ancestral allele frequency profiles β

Since β is a table of probability distributions, the values of its elements are constrained by the equality $\sum_{l=1}^{L_i}\beta_{i,l}^{k} = 1$ for all combinations of $\{i,k\}$. So to find the optimal values of β satisfying this constraint while maximizing the variational lower bound, we introduce Lagrange multipliers $\nu_{i,k}$. The new objective function to maximize is then given by:

$$
L_{new}(\mathbb{H},\mathbb{V}) = L(\mathbb{H},\mathbb{V}) + \sum_{i=1}^{I}\sum_{k=1}^{K}\nu_{i,k}\left(\sum_{l=1}^{L_i}\beta_{i,l}^{k} - 1\right) \tag{3.14}
$$

Maximizing this objective function gives:

$$\beta_{i,l}^{k} = \frac{\sum_{n=1}^{N} \xi_{n,i,l} \rho_{n,i,k}}{\sum_{l=1}^{L_i} \sum_{n=1}^{N} \xi_{n,i,l}^{n} \rho_{n,i,k}} \tag{3.15}$$

We use a uniform Dirichlet prior with hyperparameter λ (which is fixed) on each multinomial $\vec{\beta}_i^k$. Under this prior, it is not difficult to show that the estimate of $\beta_{i,l}^k$ changes to

$$\beta_{i,l}^{k} = \frac{\lambda + \sum_{n=1}^{N} \xi_{i,l}^{n} \rho_{i,k}^{n}}{\lambda * L_i + \sum_{l=1}^{L_i} \sum_{n=1}^{N} \xi_{i,l}^{n} \rho_{i,k}^{n}} \tag{3.16}$$

3.4.3 Estimating the Dirichlet prior on populations α

For estimating α we use the method described by Minka in [Minka, 2000]. This gives a Newton-Raphson iteration for α that does not involve inversion of the Hessian, and hence is reasonably fast. The log-likelihood terms involving α are:

$$L(\mathbb{H}, \mathbb{V}) = N \log \Gamma \left(\sum_{k=1}^{K} \alpha_k \right) - N \sum_{k=1}^{K} \log \Gamma \left(\alpha_k \right) + \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\alpha_k - 1 \right) \left[\psi \left(\gamma_{n,k} \right) - \psi \left(\sum_{k=1}^{K} \gamma_{n,k} \right) \right] \tag{3.17}$$

The gradient of the log-likelihood with respect to α_k is given by

$$g_k = \frac{dL(\mathbb{H}, \mathbb{V})}{d\alpha_k} = N\psi \left(\sum_{k=1}^{K} \alpha_k \right) - N\psi(\alpha_k) + \sum_{n=1}^{N} \left[\psi \left(\gamma_{n,k} \right) - \psi \left(\sum_{k=1}^{K} \gamma_{n,k} \right) \right] \tag{3.18}$$

The second derivatives, which form the Hessian, can be computed as:

$$\frac{dL(\mathbb{H}, \mathbb{V})}{d^2\alpha_k} = N\psi' \left(\sum_{k=1}^{K} \alpha_k \right) - N\psi'(\alpha_k) \tag{3.19}$$

$$\frac{dL(\mathbb{H}, \mathbb{V})}{d\alpha_k \alpha_j} = N\psi' \left(\sum_{k=1}^{K} \alpha_k \right) \qquad (k \neq j) \tag{3.20}$$

where ψ', the trigamma function, is the derivative of the digamma function. The Hessian can then be written as:

$$\mathbf{H} = \mathbf{Q} + \mathbf{1}\mathbf{1}^{\mathrm{T}} z \tag{3.21}$$

$$q_{j,k} = -N\psi^{'} \left(\alpha_k \right) \delta \left(j - k \right) \tag{3.22}$$

$$z = N\psi^{'} \left(\sum_{k=1}^{K} \alpha_k \right) \tag{3.23}$$

where $\mathbf{Q}$ is a K$\times$K matrix with elements $q_{j,k}$. As we can see from the definition, $\mathbf{Q}$ is a diagonal matrix. The Newton update equation we have is:

$$\alpha^{\text{new}} = \alpha^{\text{old}} - \left(\mathbf{H^{-1}g}\right) \tag{3.24}$$

The inverse of the Hessian can be computed using the Sherman-Morrison formula to be

$$\mathbf{H}^{-1} = \mathbf{Q}^{-1} - \frac{\mathbf{Q}^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{Q}^{-1}}{1/z + \mathbf{1}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{1}} \tag{3.25}$$

Therefore, we have that the update term is:

$$(\mathbf{H}^{-1}\mathbf{g})_k = \frac{g_k - b}{q_{k,k}} \tag{3.26}$$

where

$$b = \frac{\sum_{k=1}^{K} g_k/q_{k,k}}{1/z + \sum_{k=1}^{K} 1/q_{k,k}}$$

So the update equation for α_k is

$$\alpha_k^{\text{new}} = \alpha_k^{\text{old}} - \frac{g_k - b}{q_{k,k}} \tag{3.27}$$

3.4.4 Estimating the ancestral alleles μ and the mutation parameters δ

It is straightforward to derive gradient updates for the ancestral alleles μ and the mutation parameter δ and the details can be obtained in [Shringarpure and Xing, 2009]. While the gradient methods developed are useful for small datasets, they are inefficient on larger datasets and increase the time required for estimation. Hence we develop a couple of approximations that help speed up the hyperparameter estimation. A careful look at the results that have been produced indicates that once the founder alleles have been picked initially by fitting a mixture of mutation distributions individually at each locus, the later gradient descent on μ only makes very minor changes in their values, if any at all. Thus, to improve the speed of the algorithm, we do not perform gradient descent on the founder alleles μ but fix them after initialization. We show below an approximation for estimating the mutation parameter δ.

For the estimation of the mutation parameter δ, the only relevant term in the likelihood lower bound is the term:

$$\begin{aligned} L(\delta_i^k) = &\sum_{n=1}^{N}\sum_{l=1}^{L_i} \xi_{n,i,l}\rho_{n,i,k} \times \log f(x_{n,i};\mu_{i,l},\delta_i^k) \\ &+ \frac{(\delta_i^k)^{\zeta_1-1}\left(1-\delta_i^k\right)^{\zeta_2-1}}{B\left(\zeta_1,\zeta_2\right)} \\ &+ (\text{Terms not involving } \delta_i^k) \end{aligned} \tag{3.28}$$

where we use $\beta(\zeta_1, \zeta_2)$ as a beta prior on the mutation parameter δ. This constrains the mutation parameter to allow meaningful interpretation by using a β prior with a small expected value (around 0.1). For the mutation distribution, we use the discrete distribution whose pdf is

$$f(x|\mu,\delta) = \frac{(1-\delta)\delta^{|x-\mu|}}{1+\delta-\delta^{\mu}} \tag{3.29}$$

Approximation

We will assume δ to be small in Equation 3.29. So we can ignore the term exponential in μ in the denominator, reducing it to only $(1+\delta)$. The expansion of $(1+\delta)^{-1}$ is given by

$$\frac{1}{1+\delta} = 1-\delta+\delta^2-\delta^3+\ldots \tag{3.30}$$

$$\geq 1-\delta \tag{3.31}$$

This gives us a lower bound to the mutation distribution to be

$$f_{lb}(x|\mu,\delta) = (1-\delta)^2\delta^{|x-\mu|} \tag{3.32}$$

It is not hard to show that using this form for the mutation distribution allows a closed-form MLE for δ. This approximation gives us a lower bound to the likelihood that is not as tight as the variational lower bound. However, it offers a significant improvement in time complexity due to the existence of a closed form solution, thus avoiding the need for slow gradient-based methods. Under this approximation, the maximum-likelihood estimate of δ_i^k for the microsatellite mutation model is given by

$$\delta_i^k = \frac{\zeta_1 + \sum_{n=1}^{N}\sum_{l=1}^{L_i}\xi_{n,i,l}\rho_{n,i,k}|x_{n,i}-\mu_{i,l}|}{\zeta_2 + \sum_{n=1}^{N}\sum_{l=1}^{L_i}\xi_{n,i,l}\rho_{n,i,k}(2+|x_{n,i}-\mu_{i,l}|)} \tag{3.33}$$

3.5 Experiments and Results

We validated our model on synthetic microsatellite datasets simulated using a coalescent model to assess the performance of *mStruct* in terms of the accuracy and consistency of the estimated structure vectors, and to test the correctness of the inference and estimation algorithms we developed. We also conduct empirical analysis using *mStruct* of two real datasets: the HGDP-CEPH cell line panel of microsatellite loci and the HGDP SNP data, in comparison with the *Structure* program (version 2.2).

3.5.1 Validations on Coalescent Simulations

To verify the correctness of the empirical admixture estimations based on *mStruct* when the truth is known, we first simulated a multitude of admixture population data sets, using coalescent techniques described in [Hudson, 1990], under various user-specified admixing scenarios. Specifically, following Hudson (personal communications), without loss of generality we simulated genealogy trees for two discrete populations of effective size 2N, which were assumed to have split from a single ancestral population, also of size 2N, at a time N generations in the past. We assumed that there was no migration between the populations after the split. These two discrete populations were joined together to form a single random mating population. (A simulation of multiple-population admixing is possible, but tedious, and thus omitted here for simplicity.) After a single generation of random mating, samples were collected from the resulting population. Individuals, therefore, have i parents from population 1, and $2 - i$ parents from population 2 with probability $\binom{2}{i}/4$. Every locus was simulated independently. Microsatellite mutation was modeled by a simple stepwise mutation process. The mutation parameter $4N\mu$ was varied over data points, with 3 discrete values , $\{8, 16, 32\}$, being used. Since the expected number of mutations within the populations is given by $2N\mu$, the values chosen are representative of the diversity observed in real data [Pritchard et al., 2000a].

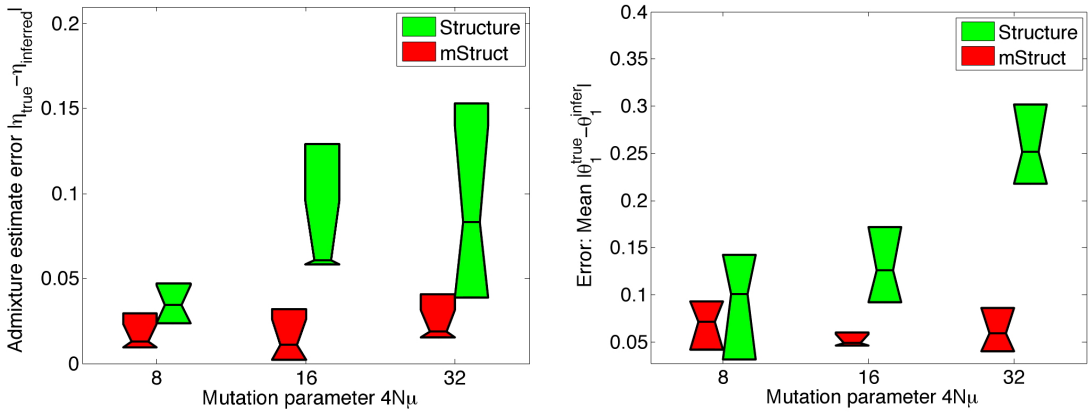
For each individual, we stored the fractional contribution of population 1 to its genome. For

each data set, we also stored the fractional contribution of population 1 to the entire population. To ensure that each population was well-represented in the admixed population, only datasets which had roughly equal contribution from both populations were accepted (The contribution of population 1 to the resulting population was required to be in $[50-0.01, 50+0.01]$ percent). For each data point in the graph, 10 data sets were simulated using the same parameter settings for the mutation parameter. Each data set had 60 individuals from the admixed population measured at 100 loci. For each data set, 10 runs of each software (i.e., *mStruct* and *Structure*) were used to determine the run with best likelihood. The statistics used in the result were computed only on the run with the best likelihood.

We use the simulated data sets to carry out three analyses. Firstly, we study the ability of both softwares to recover the contribution of population 1 (denoted as η) to the resulting admixed population. Next, we study how well each software is able to recover the proportion of ancestry in population 1 for each individual. Finally, we consider the problem of model selection- i.e., choosing the number of ancestral populations to provide an appropriate representation of the data.

Recovering the contribution of population 1 to the resulting population: We evaluated the accuracy of the estimated η under three different conditions, one for each value of the magnitude of the mutation parameter described above. The greater the magnitude, the more difficult the estimation of admixing coefficient η, because more discrepancy would exist between the ancestral alleles and the simulated population alleles. As a measure of error, we used the absolute difference between the true value η^{true} and the inferred value η^{infer}. The results shown in Figure 3.3(a) denote the means and quartiles of the result statistics. From the figure, we can see that as the magnitude of the mutation parameter increases, the error for *Structure* increases. However, for *mStruct*, there is no significant effect of the mutation parameter on the error. *mStruct* also performs better than *Structure* over all the data points.

Recovering the contribution of population 1 to the ancestry of an individual: We used



(a) Recovery of population 1 ancestry in the resultant population.

(b) Recovery of population 1 ancestry in each admixed indvidual

Figure 3.3: Recovery of individual and population level admixture parameters.

the same data from the earlier experiment for this analysis. In this case, we used the mean of the absolute difference between the true and inferred values of the proportion of ancestry of individuals in population 1 as the measure of error. Figure 3.3(b) show the results of this analysis. The results follow a similar trend as in the earlier experiment. For *Structure*, an increase in the mutation parameter causes as increase in the error, but there is no significant effect of the mutation parameter on the error for *mStruct*. We show the results for a particular data set with mutation parameter $4N\mu = 32$ in Figure 3.4. Figure 3.4(a) shows the true ancestry proportion map for the sample. It shows that around half the individuals are admixed. Figure 3.4(b) and 3.4(c) show the ancestry proportion maps inferred by *Structure* and *mStruct* respectively. We can see that the ancestry structure recovered by *mStruct* is very close to the true ancestry proportions. The recovery of ancestry proportions by *Structure* is not very close to the truth in this case.

Model selection - choice of K: As in *Structure*, our model is defined for a particular value of K, the number of ancestral populations. In general, it is not clear what value of K must be chosen to interpret the data appropriately. We performed an experiment on the simulated data to determine the most appropriate number of ancestral populations for the data. In this case, only a single data set was used with the mutation parameter $4N\mu$ set to 16. For each value of K from

1 to 5, we performed 10 runs of *mStruct* on the data and choose the run with the best likelihood for model selection. To choose the best value of K, we used the BIC criterion [Schwarz, 1978] (that we previously used to decide the optimal number of ancestral alleles at each locus). The preferred model is the one which has the minimum value of BIC. Table 3.5.1 shows the BIC values for the values of K. From the table, we can see that the model with $K = 2$ ancestral populations is correctly chosen as the optimal model.

K	**BIC**
1	6.91×10^4
2	$\mathbf{6.87 \times 10^4}$
3	6.99×10^4
4	7.12×10^4
5	7.26×10^4

Table 3.1: Model selection for simulated data: BIC values for K from 1 to 5. The model having the smallest BIC value ($K = 2$ in this case) is preferred.

3.5.2 Empirical Analysis of HGDP Datasets

The HGDP-CEPH cell line panel [Cann et al., 2002, Cavalli-Sforza, 2005] used in [Rosenberg et al., 2002] contains genotype information from 1056 indviduals from 52 populations at 377 autosomal microsatellite loci, along with geographical and population labels. The HGDP SNP data [Conrad et al., 2006] contains the SNPs genotypes at 2834 loci of 927 unrelated individuals that overlap with the HGDP-CEPH data. To make results for both types of data comparable, we chose the set of only those individuals present in both datasets. As in [Rosenberg et al., 2002], the choice of the total number of ancestral populations can be left to the user; we tried K ranging from 2 to 5, and we applied BIC to decide the Bayes optimal number of ancestral populations within this range to be $K = 4$. Below, we present the structural analysis under four ancestral

populations.

Structural map from the HGDP-CEPH data

We compare the structural maps inferred from the microsatellite data using *mStruct* and *Structure* in Figure 3.5. The most obvious difference between the maps produced by both programs is the degree of admixing that the individuals in the program are assigned. *Structure* assigns each geographical population to a distinct ancestral allele frequency profile. This assignment is very useful for partitioning individuals into separate clusters. However, in doing so, it is unable to capture the genetic structural relationships between individuals. It offers no insights into the admixture history of populations, as *mStruct* does. In contrast, the structure map produced by *mStruct* from microsatellite data suggests that all populations share a common ancestral population as a unique extra component (represented by the magenta color in Figure 3.5) that characterizes their particular regional genotypes. A structure map, characterized thus by an underlying commonality in a part of the genetic ancestry, together with regional differences, clearly reveals the expansion of humans out of Africa [Hammer et al., 1998, Templeton, 2002]. It is in this regard that *Structure* and *mStruct* are significantly different.

Both structure maps show that individuals having a similar population label (at regional, national or continental levels) have similar admixture proportions. The similarity is least if two individuals come from different continents, and most if two individuals are from the same region. We can therefore represent each regional population by the average of the admixture proportions of all individuals from the region. We computed the Euclidean distance between all pairs of the 52 regional populations and constructed a neighbour-joining tree from the distance matrices. Figures 3.6(a), 3.6(b) show the neighbour-joining trees constructed for *Structure* and *mStruct*. It is important to note that the distance measure used is not known to be a true measure of evolutionary distance. These trees have been constructed from a single instance of the distance matrix and have not been bootstrapped. Despite this, we can see that the *mStruct* tree agrees

quite well with previously constructed phylogenetic trees for human populations [Bowcock et al., 1994]. The phylogeny from *mStruct* appear to be more interpretable than that from *Structure*. In Figure 3.6(b), we can see a tighter cluster for the African populations and that American populations diverged after Asian and European diverged, rather than before.

Analysis of the mutation spectrums

Now we report a preliminary analysis of the evolutionary dynamics reflected by the estimated mutation spectrums of different ancestral populations (denoted "am-spectrum"), and of different modern geographical populations (denoted "gm-spectrum"), which is not possible by *Structure*. For the am-spectrum (Figure 3.7(a)), we compute the mean mutation rates over all loci and founding alleles for each ancestral population as estimated by *mStruct*. We estimate the gm-spectrum (Figure 3.7(b)) as follows: for every individual, a mutation parameter is computed as the per-locus number of observed alleles that are attributed to mutations, weighted by the mutation parameters corresponding to the ancestral allele chosen for that locus. This can be computed by observing the population-indicator (Z) and the allele-indicator (C) for each locus of the individual. We then compute the population mutation parameters by averaging mutation parameters of all individuals having the same geographical label.

As shown in the gm-spectrum in Figure 3.7(b), the mutation parameters for African populations are indeed higher than those of other modern populations. Since the mutation parameter reflects effects of mutation rate and population age, this indicates that they diverged earlier, a common hypothesis of human migration. Other trends in the gm-spectrums also reveal interesting insights. We computed the empirical mutation parameters for each of the 52 subpopulations present in the data as we did for each continent. Since each population has an associated latitude and longitude, this allows us to set up a function that maps a geographical latitude/longitude coordinate to an empirical mutation parameter. Figure 3.8 shows the contour plot of this function. The mutation parameter δ in our model is a measure of variability (a combination of per genera-

tion mutation rate and age of the population). Thus, the contour plots shows us how the amount of variability changes across the world. We can see that the maximum variation is in Africa. There is a decrease in variation as we move away from central Africa. We can also see that the South American tribes have the least amount of accumulated variation. This is in qualitative agreement with the ages of different populations as predicted by the "Out of Africa" hypothesis of human migration.

Structural map from the HGDP SNP data

Figure 3.9 shows the structural maps produced by *mStruct* and *Structure* for the HGDP SNP data. We can see that the two population maps are nearly identical, which signals an inconsistency between the microsatellite and SNP *mStruct* results for the human data. However, there are some important caveats that must be taken into consideration. In our analysis, we consider a simplistic bernoulli-like model of SNP mutation. While richer mutation models could potentially reduce this difficulty, there is a more significant difficulty with the analysis of SNP data. The bi-allelic nature of SNP markers makes it difficult to draw any inferences about the correct number of ancestral alleles at a locus. For microsatellites, this problem is considerably easier due to their multi-allelic nature. As a result, *mStruct* is unable to obtain more information about evolutionary history from SNP markers than *Structure* does. As we have explained earlier, *mStruct* is an extension of *Structure* that finds signals about mutations present in the data. So in the event that *mStruct* is unable to find any extra mutation information from the data, it is quite reasonable to expect its output to be nearly the same as that of *Structure*.

Model selection

As with all probabilistic models, we face a tradeoff between model complexity and the log-likelihood value that the model achieves. In our case, complexity is controlled by the number of ancestral populations we pick, K. Unlike non-parametric or infinite dimensional models

(Dirichlet processes etc.), for models of fixed dimension, it is not clear in general as to what value of K gives us the best balance between model complexity and log-likelihood. In such cases, different information criteria are often used to determine the optimal model complexity. To determine what number of ancestral populations fit the HGDP SNP and microsatellite data best, we computed BIC scores for K=2 to K=5 for both kinds of data separately. The results are shown in Figure 3.10. From the BIC curves for both SNP and microsatellite data, we can see that the curves suggest K=4 as the best fit for the data.

3.6 Discussion

The task of estimating the genetic contributions of ancestral populations, i.e., structural map estimation, in each modern individual, is an important problem in population genetics. Due to the relatively high rates of mutation in markers such as microsatellites and SNPs, multilocus genotype data usually harbor a large amount of variation, which allows differentiation even between populations that have close evolutionary relationships. However, to our knowledge, none of the existing methods is able to take advantage of this property to compare how marker mutation rates vary with population and locus, while at the same time exploiting such information for population structural estimation. Traditionally, population structure estimation and mutation spectrum estimation have been performed as separate tasks.

We have developed *mStruct*, which allows estimation of genetic contributions of ancestral populations in each modern individual in light of both population admixture and allele mutation. The variational inference algorithm that we developed allows tractable approximate inference on the model. The ancestral proportions of each individual enable representing population structure in a way that is both visually easy to interpret, as well as amenable to further computational analysis.

The statistical modeling differences between *mStruct* and *Structure* provide an interesting insight into the possible reasons which lead to *mStruct* inferring higher levels of admixture than

Structure. In Structure's representation of population, every microsatellite allele is considered to be a separate element of the population, even though they might be very similar. In the inheritance model representation, such alleles are considered to be possibly derived from a single ancestral allele. This can lead to detection of extra similarity among individuals possessing these alleles. This is probably the main reason that the inferred levels of admixture are higher in *mStruct* than *Structure*.

Another parameter that would also affect inferred levels of admixture is the δ parameter which determines the variance of the mutation distributions. Higher values of δ (tending to 1) lead to significantly higher levels of inferred admixture. If a strong prior is not used, the δ values tend towards 1 in the initial few steps of the variational EM algorithm. This seems to happen due to the initial imprecise assignments for the z and c indicator variables. However, the region of high δ values is a region of low log-likelihood in the parameter space and the EM quickly finds a local optimum which is undesirable due to the low log-likelihood of that region of the parameter space.

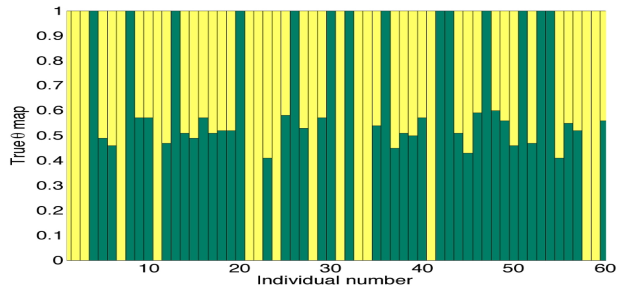
In conjunction with geographical location, the inferred ancestry proportions could be used to detect migrations, sub-populations etc. Moreover, the ability to estimate population and locus specific mutation parameters also allows us to substantiate evolutionary dynamics claims based on high/low mutation parameters in certain geographical population, or on high/low mutation parameters at certain loci in the genome. While the estimates of mutation parameters that *mStruct* provides are not on an absolute scale, the comparison of their relative magnitudes is certainly informative.

The mutation model we currently use is a computationally simple one. However, it lacks the ability to distinguish between the effects of per generation mutation rate and the age of the population. Under the Stepwise Mutation Model, we can model inheritance by using a more complex but powerful model using Bessel functions [Felsenstein and Others, 2004]. This form would allow separate inference of the per generation mutation rate as well as the age of the

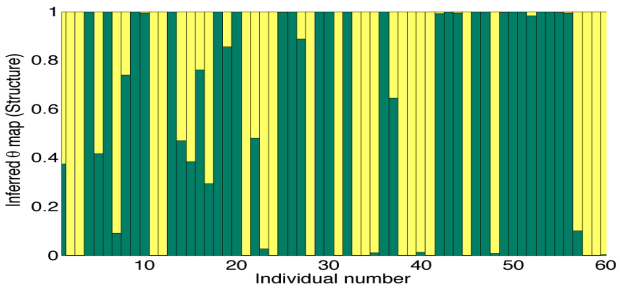
population.

As of now, there remain a number of possible extensions to the methodology we presented so far. It would be instructive to see the impact of allowing linked loci as in [Falush et al., 2003]. We have not yet addressed the issue of the most suitable choice of mutation process, but instead have chosen one that is reasonable and computationally tractable. It would also be interesting to combine *mStruct* with the nonparametric Bayesian models based on the Dirichlet processes as in programs such as Spectrum [Sohn and Xing, 2007] and Structurama [Huelsenbeck and Andolfatto, 2007].

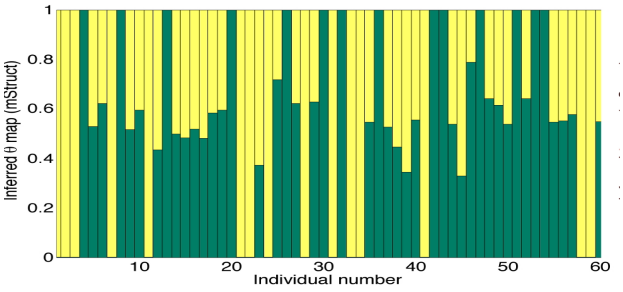
In summary, current population stratification methods such as *Structure* ignore the effects of allele mutations, which are a significant factor in shaping allele diversity in microsatellites in human populations. In doing so, they are restricted to clustering human genetic data rather than being able to identify admixing of populations. Clustering is useful for population stratification, but a more accurate representation of events such as genome variations might cast more light on population evolutionary history. By incorporating the effect of allele mutations, the *mStruct* approach developed in this paper represents such an attempt to gain more insight into the fine structures of genetic admixing of populations and their divergence times.



(a)



(b)



(c)

Figure 3.4: A comparison of the true and inferred ancestry proportions for a single example. (a) The true ancestry proportions for the sample. (b) The ancestry proportions inferred by *Structure*. (c) The ancestry proportions inferred by *mStruct*.

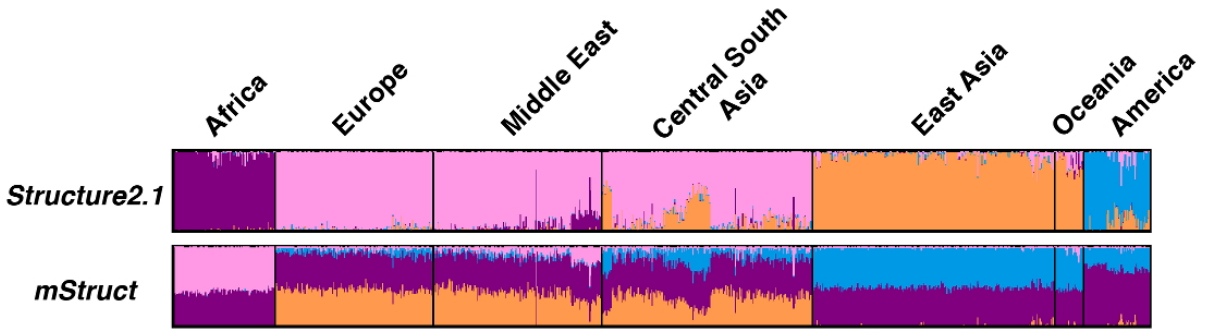
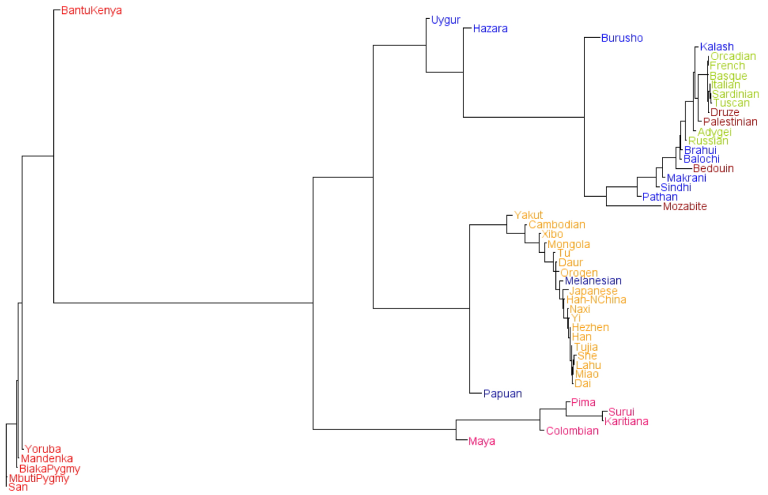
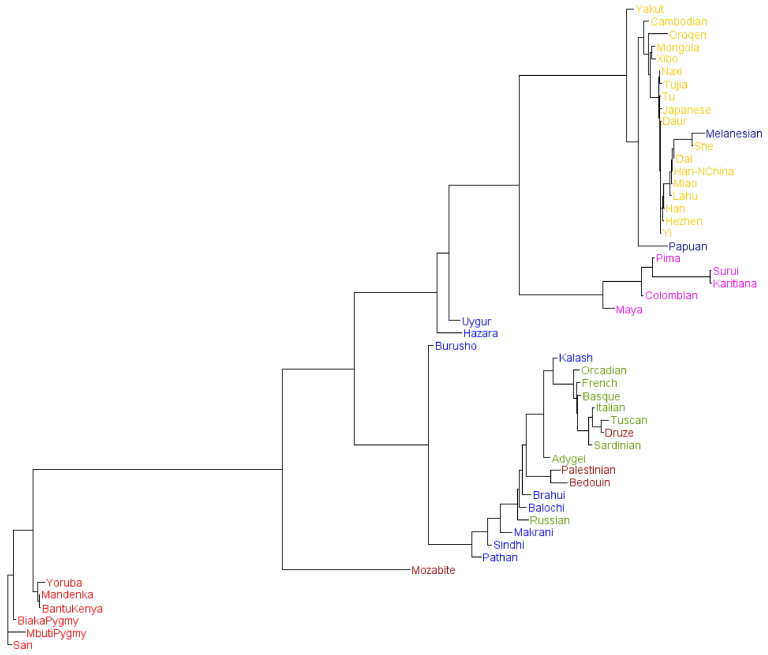


Figure 3.5: Ancestry structure maps inferred from microsatellite portion of the HGDP dataset, using *mStruct* and *Structure* with 4 ancestral population. The colors represent different ancestral populations.

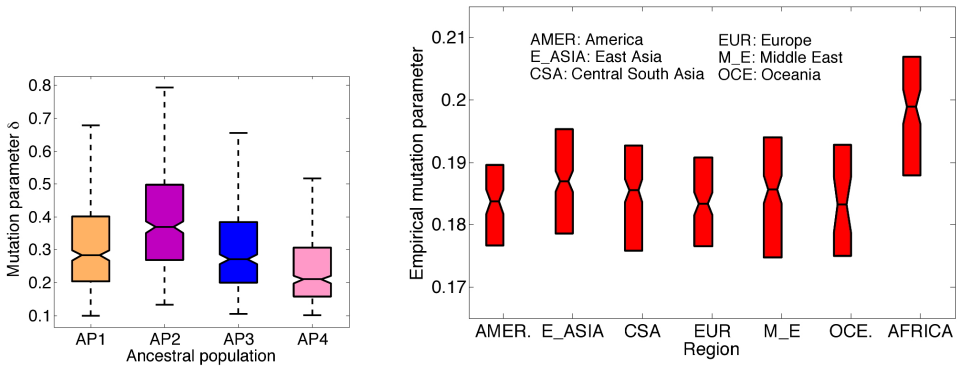


(a) *Structure* tree



(b) *mStruct* tree

Figure 3.6: Neighbour-joining trees constructed using *mStruct* and *Structure* for the 52 regional populations in the HGDP microsatellite data



(a) Am-spectrum (b) Gm-spectrum

Figure 3.7: Am-spectrum and Gm-spectrum inferred from microsatellite portion of the HGDP dataset, using *mStruct* with 4 ancestral population. The colors represent different ancestral populations.

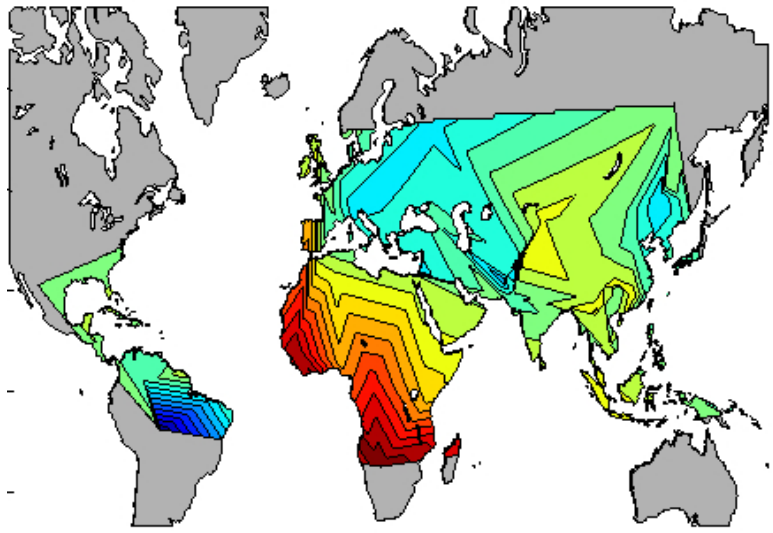


Figure 3.8: Contour map of the empirical mutation parameters over the world map

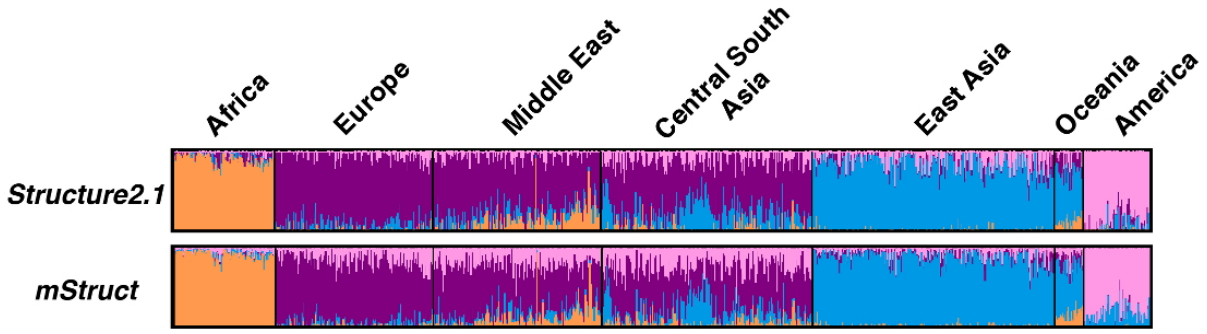


Figure 3.9: Ancestry structure maps inferred from SNPs portion of the HGDP dataset, using *mStruct* and *Structure* with 4 ancestral population.

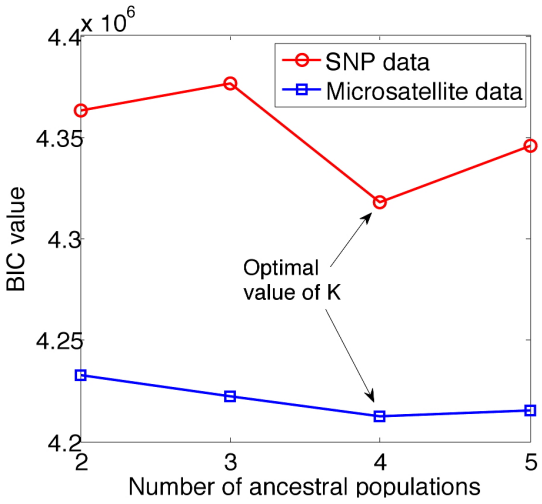


Figure 3.10: Model selection with BIC score for the HGDP data with *mStruct* on SNP and microsatellite data

Chapter 4

How many ancestral populations? A nonparametric Bayesian approach

Clustering of genotype data is an important way of understanding similarities and differences between populations. A summary of populations through clustering allows us to make inferences about the evolutionary history of the populations. Many methods have been proposed to perform clustering on multi-locus genotype data. However, most of these methods do not directly address the question of how many clusters the data should be divided into and leave that choice to the user.

We present *StructHDP* [Shringarpure et al., 2011], which is a method for automatically inferring the number of clusters from genotype data in the presence of admixture. Our method is an extension of two existing probabilistic clustering methods, *Structure* and *Structurama*. Using a Hierarchical Dirichlet Process, we model the presence of admixture of an unknown number of ancestral populations in a given sample of genotype data. We use a Gibbs sampler to perform inference on the resulting model and infer the ancestry proportions and the number of clusters that best explain the data.

To demonstrate our method, we simulated data from an island model using the neutral coalescent. Comparing the results of *StructHDP* with *Structurama* shows the utility of combining

HDPs with the *Structure* model. We used *StructHDP* to analyze a data set of 155 Taita thrush, *Turdus helleri*, which has been previously analyzed using *Structure* and *Structurama*. *StructHDP* correctly picks the optimal number of populations to cluster the data. The clustering based on the inferred ancestry proportions also agrees with that inferred using *Structure* for the optimal number of populations. We also analyzed data from 1048 individuals from the Human Genome Diversity project from 53 world populations. We found that the clusters obtained correspond with major geographical divisions of the world, which is in agreement with previous analyses of the dataset.

4.1 Introduction

An important question that needs to be addressed when solving the problem of population stratification is deciding how many populations are required to best explain the variation observed in a given set of individuals. The Bayesian models described in Sections 2.3 and 3.2 require the user to specify the number of clusters (ancestral populations) into which the individuals are divided. However, this might not always be possible or desirable, in the absence of prior knowledge about the evolutionary history of the sample. A common solution to this problem is to use fixed-dimensionality models in combination with an information criterion [Akaike, 1974, Gao et al., 2011, Schwarz, 1978] to decide the number of ancestral populations. To address this problem, an extension of *Structure* was developed by Pella and Masuda [2006] using Dirichlet processes [Ferguson, 1973]. Based on their method, Huelsenbeck and Andolfatto [2007] developed *Structurama*. *Structurama* automatically infers the number of population clusters into which a given data set should be divided provided individuals only belong to a single population. Coalescent simulations by Huelsenbeck and Andolfatto [2007] using island models show that inference of the number of populations is accurate when migration rates are low and differentiation between populations is high. However, the assumption that each individual only belongs to a single ancestral population implies that *Structurama* is unable to model admixture between

ancestral populations.

We develop *StructHDP*, a method for automatically inferring the number of population clusters present in a group of individuals, while accounting for admixture between ancestral populations. Using the Hierarchical Dirichlet Process framework for clustering developed by Teh et al. [2005], we extend the *Structure* model so that the number of populations is inferred by the model and need not be specified by the user. This work is also an extension of the Dirichlet process model developed by Pella and Masuda [2006] which has been implemented in *Structurama*.

We simulated data from an island model using the neutral coalescent to test the performance of our method at recovering the true number of ancestral populations. Comparing the results of *StructHDP* with *Structurama* shows the utility of combining HDPs with the *Structure* model. We used *StructHDP* to analyze a set of 155 Taita thrush individuals, *Turdus helleri*. This dataset has been previously analyzed using *Structure* and *Structurama*. We found that *StructHDP* correctly identifies the optimal number of populations to cluster the data. The clustering enforced by the inferred ancestry proportions for individuals also agrees with that inferred using *Structure* with the appropriate choice of the number of populations K. We also analyzed a set of 1048 individuals from the Human Genome Diversity Project (HGDP) using *StructHDP*. We found that the clusters inferred coincide with the major geographical divisions present in the data. We also observed that the distance between populations (based on their cluster memberships) is strongly positively correlated with F_{st} between populations, which suggests that the inferred cluster memberships capture the genetic variation present in the data well.

4.2 Related work

4.3 Approach

We approach the problem of finding the number of ancestral populations by extending the admixture model of *Structure* to a setting where there are potentially infinite ancestral population

components in the mixture. Performing inference then allows us to examine the number of ancestral populations that have a non-zero contribution to the set of individuals under consideration. We use the Hierarchical Dirichlet Process framework [Teh et al., 2005] to model the mixture of infinite ancestral populations.

Consider the problem of clustering the markers within a single individual based on their population of origin. We can assume that the number of populations that contribute to the single individual's genome is unknown and is a random variable. The Dirichlet process (DP) [Ferguson, 1973] was proposed to solve a problem of this nature, where objects (genetic markers) belong to one of a potentially infinite number of mixture components (ancestral populations). In the case of multiple individuals, we can posit multiple DPs, one for each individual, that will address the problem of not knowing the optimal number of populations. We also require that the ancestral populations inferred for the DPs be the same across all the individuals. Mathematically, this is analogous to ensuring that mixture components are shared across DPs.

The Hierarchical Dirichlet process (HDP) is a framework for clustering of observations when the observations are present in groups. Each group can be modeled using a finite mixture model or a Dirichlet process. The mixture models or DPs across groups are linked by sharing mixture components. It is useful to think of each group as having its own Dirichlet processes, with the processes linked to each other by the parameters of the HDP. *StructHDP* is based on the Hierarchical Dirichlet process described by Teh et al. [2005]. In the following section, we provide a description and mathematical representation of the HDP model.

4.4 Model

A commonly used analogy for representing HDPs is the Chinese Restaurant Franchise (CRF). This is an extension of the representation of the Dirichlet process (DP) as a Chinese restaurant with customers. The DP representation and its application to *Structurama* are described in more detail by Huelsenbeck and Andolfatto [2007]. A CRF comprises of a number of Chinese

restaurants which share a common (possibly infinite) menu of dishes. In a CRF, each restaurant corresponds to a group of observations, and the customers are observations. The dishes served in the restaurant are the mixture components, and sharing of mixture components across groups corresponds to sharing of dishes across restaurants. In the CRF metaphor, a new customer (observation) arrives at the restaurant corresponding to its group. The customer chooses a previous occupied table in the restaurant with a probability proportional to the number of customers already at the table, or, with a constant probability, chooses a new table. Every table serves a dish from the possible set of dishes, and every customer at the table is assigned that particular dish, i.e, the observation is assigned the particular mixture component that is associated with the table. All observations that are assigned to a particular table are considered to originate from the same mixture component, clustering the observations within the group. The same mixture component might also be shared across multiple tables within a group. The method of choosing a table for a new customer is similar to a "rich gets richer" model which is regulated by the probability of starting a new table. This is the property of the HDP that is responsible for its clustering behavior.

This analogy can be easily extended to the case of genetic data, with every individual considered to be a separate group corresponding to a restaurant. The loci within an individual are the customers in the restaurant, and the ancestral populations are the mixture components or the dishes in the CRF. A minor subtlety that arises in this case is that the set of possible alleles at each locus might be different, which needs to be accounted for in the inference process. This can be accomplished easily with some minor additional bookkeeping without changing the inference process significantly.

Consider a dataset having N individuals genotyped at M loci. The observed allele for individual j at locus i is denoted by x_{ji}. For ease of representation, we will ignore the diploid nature of genotype data. In implementation, we shall allow our method to handle data of any fixed ploidy. The HDP can then be used to generate the allele x_{ji} for the j^{th} individual at the i^{th} locus

as follows:

$$\begin{aligned} G_0|\gamma, H_i &\sim DP(\gamma, H_i) \\ G_j|\alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \zeta_{ji}|G_j &\sim G_j \\ x_{ji}|\zeta_{ji} &\sim \text{F}(\zeta_{ji}) \end{aligned}$$

Here, H_i is the base distribution over alleles at locus i, commonly a Dirichlet distribution. γ and α_0 are parameters of the HDP that control how fast new populations are added to the model. G_0 is an intermediate probability distribution over alleles at locus i and G_j is a distribution specific to individual j. The individual-specific distributions G_j are connected to one another through G_0 and α_0, ensuring the sharing of ancestral populations across individuals. G_0 and G_j are both generated by Dirichlet processes (DP) that use γ and α_0 as parameters. The ζs denotes the mixture components. x_{ji} is a sample from a distribution $F(\zeta_{ji})$, a multinomial distribution over alleles in our case.

For modeling purposes, it is helpful to modify the representation of the HDP so that the generative process looks as follows.

$$\beta|\gamma \sim \text{GEM}(\gamma) \tag{4.1}$$

$$\pi_j|\alpha_0, \beta \sim \text{DP}(\alpha_0, \beta) \tag{4.2}$$

$$\phi_{ik}|H_i \sim H_i \tag{4.3}$$

$$z_{ji}|\theta_j \sim \text{Multinomial}(1, \theta_j) \tag{4.4}$$

$$x_{ji}|z_{ji}, (\phi_{ik})_{k=1}^{\infty} \sim \text{F}(\phi_{z_{ji}}) \tag{4.5}$$

where we say that $\beta = (\beta_k)_{k=1}^{\infty} \sim \text{GEM}(\gamma)$ if it satisfies the following construction:

$$\beta'_k|\gamma \sim \text{Beta}(1, \gamma) \tag{4.6}$$

$$\beta_k = \beta'_k \prod_{l=1}^{k-1} (1 - \beta'_l) \tag{4.7}$$

This construction ensures that $\sum_{k=1}^{\infty} \beta_k = 1$. The β thus represents the fractional contributions of the potentially infinite populations to the given set of individuals.

In the HDP representation above, ϕ_{ik} represents the allele frequencies of the k^{th} population at the i^{th} locus. θ_j is a vector that denotes the ancestry proportions (contributions from all populations) for individual j, and its components sum to 1. The indicator variable z_{ji} denotes which population the observed allele x_{ji} at locus i originates from. We will use this notation for representing the HDP model for our problem due to its similarity with the *Structure* generative process. This representation also shows how the model can account for diploid individuals by changing the step of sampling z_{ji} and x_{ji} to the following:

$$
\begin{aligned}
z_{ji,1}|\theta_j &\sim \text{Multinomial}(1, \theta_j) \\
z_{ji,2}|\theta_j &\sim \text{Multinomial}(1, \theta_j) \\
x_{ji,1}|z_{ji,1}, (\phi_{ik})_{k=1}^{\infty} &\sim \text{F}(\phi_{z_{ji,1}}) \\
x_{ji,2}|z_{ji,2}, (\phi_{ik})_{k=1}^{\infty} &\sim \text{F}(\phi_{z_{ji,2}})
\end{aligned}
$$

where $x_{ji,1}$ and $x_{ji,2}$ now represent the two alleles at locus i in individual j and $z_{ji,1}$ and $z_{ji,2}$ are their respective population indicator variables. This allows the model to account for mixed ancestries at a single locus as well. For ease of representation, we will drop the subscript indicating the ploidy in the analysis.

Figure 4.1 shows the graphical model representation of the *StructHDP* generative process. In this graphical model representation, the nodes represent random variables which have been described earlier. The edges denote dependencies between the random variables due to the sampling steps in the generative process. The shaded nodes represent the random variables we observe, viz, the alleles observed at each locus.

To allow for more flexibility with the parameter settings, we impose priors on α_0, γ and the base distributions H_i. We assume that α_0 and γ have Gamma priors with parameters (α_a, α_b) and (γ_a, γ_b) respectively and that H_i has a symmetric Dirichlet distribution with parameter λ.

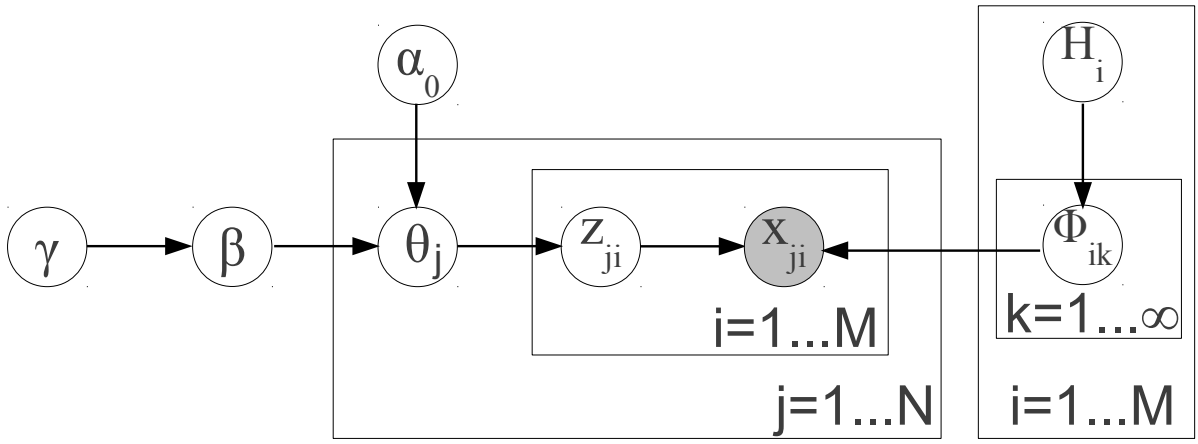


Figure 4.1: Graphical model representation of the generative process of *StructHDP*. Nodes represent random variables and edges indicate dependencies between random variables. The shaded circle indicates the observed alleles. The dataset has N individuals, each genotyped at M loci. For ease of representation, we do not show the ploidy of the individual in the graphical model.

$$\alpha_0 \sim \text{Gamma}(\alpha_a, \alpha_b) \tag{4.8}$$

$$\gamma \sim \text{Gamma}(\gamma_a, \gamma_b) \tag{4.9}$$

$$H_i \sim \text{Dir}(\lambda) \tag{4.10}$$

4.4.1 Inference

For performing inference on the model, we use Gibbs sampling, an MCMC sampling method, described for the HDP by Teh et al. [2005]. For inference in the CRF representation of the HDP, we create some bookkeeping variables **m** that keep count of the number of tables at the restaurant and franchise levels.

Inference steps

Using all the variable updates, the inference process can be described as:

1. Set the values for the prior parameters $\alpha_a, \alpha_b, \gamma_a, \gamma_b, \lambda$.

2. Start with random values for all other variables.

3. Sample $\mathbf{z}$ variables given all other variables.

4. Sample $\mathbf{m}$ variables given all other variables, using updated value of $\mathbf{z}$.

5. Sample β given all other variables, using updated values of $\mathbf{z}$ and $\mathbf{m}$.

6. Sample α_0 using updated values of $\mathbf{z}$, $\mathbf{m}$ and β.

7. Sample γ using updated values of all other variables.

8. Repeat 3-7 until convergence.

The Gibbs sampling update distributions can be derived following the methodology in Teh et al. [2005]. We describe the details of the Gibbs sampling update distributions and their derivations below.

The population allele frequencies at locus i are assumed to be $\{\phi_{i1}, \cdots, \phi_{iK}\}$ where K can be infinity and only a finite number of the populations are used in the dataset. The prior over the allele frequencies ϕ_{ik} is H_i. In the restaurant analogy, we use t_{ji} to denote the table for customer x_{ji}, n_{jtk} to denote the number of customers in restaurant j at table t eating dish k, while m_{jk} denotes the number of tables in restaurant j serving dish k. Marginal counts are represented with dots. So $n_{jt.}$ denotes the number of customers in restaurant j at table t, and $m_{..}$ represents the total number of tables in the franchise.

Let $\mathbf{x} = (x_{ji} : \text{all } j, i), \mathbf{x}_{jt} = (x_{ji} : \text{all } i \text{ with } t_{ji} = t), \mathbf{t} = (t_{ji} : \text{all } j, i), \mathbf{z} = (z_{ji} : \text{all } j, i), \mathbf{m} = (m_{jk} : \text{all } j, k)$. When a superscript is used with a set of variables, e.g., x^{-ji} or $n_{jt.}^{-ji}$, this means that the variable corresponding to the index is removed from the set. In the example, $x^{-ji} = \mathbf{x}/x_{ji}$ and $n_{jt.}^{-ji}$ is the number of observations in group j associated with table t leaving out observation x_{ji}.

An important quantity we will use often in sampling is the conditional density of x_{ji} under mixture component k given all data except x_{ji}. This can be computed as

$$f_k^{-x_{ji}}(x_{ji}) = \frac{\int f(x_{ji}|\phi_{ik}) \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'|\phi_{ik}}) h(\phi_{ik}) d\phi_{ik}}{\int \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'|\phi_{ik}}) h(\phi_{ik}) d\phi_{ik}} \tag{4.11}$$

Here, we are marginalizing out the effects of the uncertainty in the allele frequencies ϕ_{ik}. For our purposes, $f(.|\theta)$ is a multinomial distribution and $h_i(.)$ is a symmetric Dirichlet distribution with parameters λ, on the simplex of dimension P (where P is the number of alleles observed at locus i). Therefore the numerator and denominator are the normalization constants of the posterior Dirichlet distributions.

At locus i, we can represent the observed alleles as $\{a_1, \cdots, a_P\}$. Then we have that

$$f(x_{ji}|\phi_{ik}) = \prod_p \phi_{ik,p}^{\mathcal{I}[x_{ji}=a_p]} \tag{4.12}$$

Using this in Equation 4.11 gives us,

$$f_k^{-x_{ji}}(x_{ji}) = \frac{B(h_1 + \sum_{j'i', z_{j'i'}=k} \mathcal{I}[x_{j'i'} = a_1], \cdots)}{B(h_1 + \sum_{j'i' \neq ji, z_{j'i'}=k} \mathcal{I}[x_{j'i'} = a_1], \cdots)} \tag{4.13}$$

where $B(.)$ is the multinomial beta function, which can be written in terms of the Gamma function:

$$B(\alpha_1, \cdots, \alpha_P) = \frac{\prod_{p=1}^P \Gamma(\alpha_p)}{\Gamma(\sum_{p=1}^P \alpha_p)}$$

Sampling for the population indicator variables z is given by

$$\begin{aligned} p(z_{ji} = k|\mathbf{z}^{-ji}, \mathbf{m}, \beta) &= (n_{j.k}^{ji} + \alpha_0\beta_k) f_k^{-x_{ji}}(x_{ji}) \\ &\quad , \text{ if k is previously used} \\ &= \alpha_0\beta_u f_{k^{new}}(x_{ji}), \text{ if k is new} \end{aligned}$$

To sample m, we use a result derived in [Teh et al., 2005],

$$p(m_{jk} = m|\mathbf{z}, \mathbf{m}^{-jk}, \beta) = \frac{\Gamma(\alpha_0\beta_k)}{\Gamma(\alpha_0\beta_k + n_{j.k})} s(n_{j.k}, m)(\alpha_0\beta_k)^m$$

where $s(n, m)$ are unsigned Stirling numbers of the first kind.

Sampling for β is given by

$$(\beta_1, \cdots, \beta_k, \beta_u)|\mathbf{m}, \mathbf{k} \sim \text{Dir}(m_{.1}, \cdots, m_{.K}, \gamma)$$

Concentration parameter updates

For updating the concentration parameter α_0, we use the method described by [Teh et al., 2005], using a sampling scheme of auxiliary variables. For N individuals, define auxiliary variables, $\mathbf{w} = (w_j)_{j=1}^N$ and $\mathbf{s} = (s_j)_{j=1}^N$, where each $w_j \in [0, 1]$ and each s_j is a binary variable in $\{0, 1\}$. Then we have the following sampling scheme

$$\begin{aligned} q(\alpha_0|\mathbf{w}, \mathbf{s}) &\sim \text{Gamma}\left(a + m_{..} + 1 - \sum_j s_j, b + 1 - \sum_j \log(w_j)\right) \\ q(w_j|\alpha_0) &\sim \text{Beta}(\alpha_0 + 1, n_{j..}) \\ q(s_j|\alpha_0) &\sim \text{Binomial}(1, n_{j..}/\alpha_0/(1 + n_{j..}/\alpha_0)) \end{aligned}$$

To update α_0, we iterate these three steps until the value of α_0 converges. Convergence is usually quick and takes about 20-30 iterations.

For updating γ we use the method described in Escobar and West [1995], using an auxiliary variable η. We assume that γ has a gamma prior Gamma(a, b).

We have

$$\begin{aligned} q(\gamma|\eta, K) &\sim \pi_\eta \text{Gamma}(a + k, 1/(b - \log(\eta))) \\ &\quad + (1 - \pi_\eta)\text{Gamma}(a + K - 1, 1/(b - \log(\eta))) \end{aligned}$$

where the mixture weights are given by

$$\frac{\pi_\eta}{1 - \pi_\eta} = \frac{a + k - 1}{m_{..}(b - \log(\eta))}$$

Secondly, we have

$$q(\eta|\gamma, K) \sim \text{Beta}(\gamma + 1, m_{..})$$

Alternating these updates until the value of γ converges provides a method for updating γ.

4.4.2 Other inference details

Like all MCMC methods, the sampler is run for a large number of iterations, with some initial iterations discarded as burn-in. Samples from the posterior can then be used to estimate the

ancestry proportions π_j for each individual. The posterior distribution for the individual ancestry proportions π_j can be shown to be a Dirichlet distribution.

$$\theta_j \sim \text{Dir}\left(\cdots, \alpha_0\beta_k + \sum_{i=1}^{M} \mathcal{I}[z_{ji} = k], \cdots\right) \tag{4.14}$$

where $\mathcal{I}[.]$ denotes an indicator function. If the number of populations remains constant across iterations in the sampling (as is often seen to happen in our experiments after a large number of iterations), this estimate can be averaged over multiple samples to get a more accurate estimate of the individual ancestry proportions.

As with the Gibbs sampler used in *Structure*, our method could have problems with the identifiability of clusters, if label switching for the clusters were a frequent occurrence. In practice, we find that label switching is infrequent, and can avoided by the use of the restricted growth function (RGF) notation of Stanton and White [1986] in summarizing MCMC results.

4.5 Results

4.5.1 Coalescent simulation data

We performed coalescent simulations based on an island model similar to Huelsenbeck and Andolfatto [2007]. We used the program *ms* [Hudson, 2002] to simulate samples under a neutral coalescent model. As an initial evaluation of the performance of *StructHDP* in recovering the correct number of population clusters, we simulated data from a symmetric equilibrium island model with 4 demes of equal size, with the mutation rate $\theta = 4N_e\mu = 0.5$ and migration rate $M = 4N_em = \{1, 2, 4\}$. In each case, 100 diploid individuals were sampled with an equal number being sampled from each deme. 50 replicates were created for each parameter setting.

We analyzed the data using *StructHDP*, *Structurama* and *Admixture*. For *StructHDP*, the priors for both concentration parameters were set to (0.5,0.5) and the parameter for the Dirichlet distribution of H was set to 0.5. The *StructHDP* Gibbs sampler was run for 25,000 iterations,

Method ↓ / Migration rate →	M=1	M = 2	M =4
StructHDP	0.10	**0.01**	**0.15**
Structurama	**0.0**	-0.21	-1.31
Admixture+AIC	-1.8	-1.73	-1.65
Admixture+BIC	-2.6	-2.78	-2.62
Admixture+CV	2.5	2.63	2.71

Table 4.1: Comparison of simulation results for *StructHDP*, *Structurama* and *Admixture*. 50 replicates, consisting of 100 diploid individuals each, were sampled from a 4-deme symmetric island model, with $\theta = 0.5$ and $M = \{1, 2, 4\}$. The error in recovering the number of demes is shown, as computed by the error measure $E(E(K|X) - K_T)$.

with the first 12,500 iterations discarded as burn-in. To thin the Markov chain, samples were taken every 25 iterations. We computed the expected value of the number of populations, K, using the sampled values of K from the Gibbs sampler. The expected value of K, $E(K|X)$ can then be compared against the true value of the number of demes, $K_T = 4$, across multiple replicates, to get an error measure that is given by $E(E(K|X) - K_T)$ [Huelsenbeck and Andolfatto, 2007].

For *Structurama*, the experiments for each parameter setting were performed with different priors on the expected number of populations in [Huelsenbeck and Andolfatto, 2007]. For comparison purposes, we chose the best result, i.e, the prior setting that gave the least error. Model selection with *Admixture* can be done in three different ways by choosing either the AIC, BIC or the cross-validation error as the measure of model fit. We present results for all three measures.

Table 4.1 shows the results of the simulation. We can see that the error in recovering K is much smaller for *StructHDP* than for *Structurama* and for *Admixture*, except when the migration rate is small. The underlying assumption of the Dirichlet process model of *Structurama* is that there is no admixture and individuals only belong to a single ancestral population. As a result, in a simulation setting with less admixture due to migration, the number of recovered populations for *Structurama* is almost perfect. As the amount of admixture increases, the error in the number of

recovered populations increases. On the other hand, *StructHDP* explicitly accounts for admixture in the model. Therefore it recovers the true number of demes in the island model with low error for all parameter values. In terms of F_{st}, we can say that as the F_{st} between the demes decreases (as migration increases), the accuracy of *Structurama* drops while that of *StructHDP* is unaffected.

Admixture performs worse than both *StructHDP* and *Structurama* in recovering the true number of populations. This may be due to the small number of markers that are used in the simulation study.

4.5.2 Real data analysis

Taita thrush data:

We used our method to analyze a data set of N = 155 Taita thrush, *Turdus helleri* [Galbusera et al., 2000]. Each individual was genotyped at M = 7 microsatellite loci. Individuals were sampled at four locations in southeast Kenya [Chawia (17 individuals), Ngangao (54), Mbololo (80), and Yale (4)]. The thrush data were previously analyzed in [Huelsenbeck and Andolfatto, 2007, Pritchard et al., 2000a] so we use it to verify the correctness of *StructHDP*.

We ran *StructHDP* for 25,000 iterations, with the first 12,500 iterations as burn-in. Samples were taken every 25 iterations to thin the Markov chain. The priors for both concentration parameters were set to (0.5,0.5) and the parameter for the Dirichlet distribution of H was set to 0.5.

We find that our method converges to K=3 populations in a few thousand iterations. The posterior distribution for K is shown in Figure 4.2. From the posterior, we can see that $K = 3$ is the most likely value for K. Figure 4.3 shows a single sample for the ancestry proportions of the thrush data. The clusters agree with geographical labels well except for a few individuals. We also see that the 4 Yale individuals fall into the same cluster as the Ngangao individuals. All of these findings agree with those of Pritchard et al. [2000a] when *Structure* is initialized with

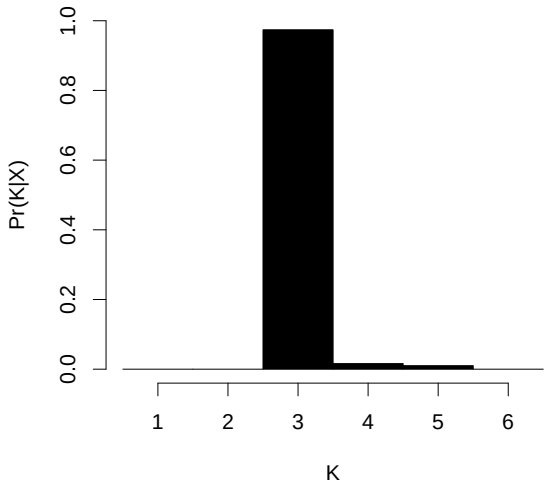


Figure 4.2: Posterior distribution for number of populations, $Pr(K|X)$ for the thrush data.

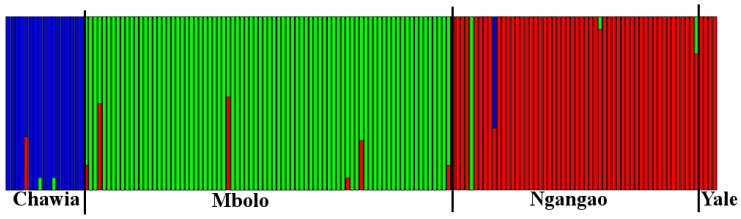


Figure 4.3: A single sample of the ancestry proportions for the thrush data. The black lines separate the individuals according to their geographic labels. The analysis did not use any geographical information.

$K = 3$ clusters. Figure 4.4 shows the results of *Structure* analysis of the thrush data with $K = 3$. In their analysis, Pritchard et al. also found that $K = 3$ explains the data best. Their conclusion was based on an *ad hoc* approximation to $Pr(K|X)$, the posterior likelihood of K given the data X, while *StructHDP* automatically infers this from the data.

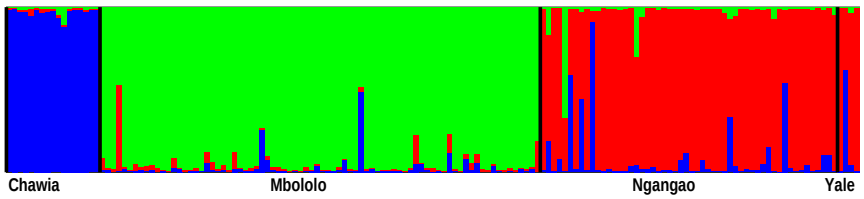


Figure 4.4: The ancestry proportions for the thrush data from a single *Structure* run for K=3.

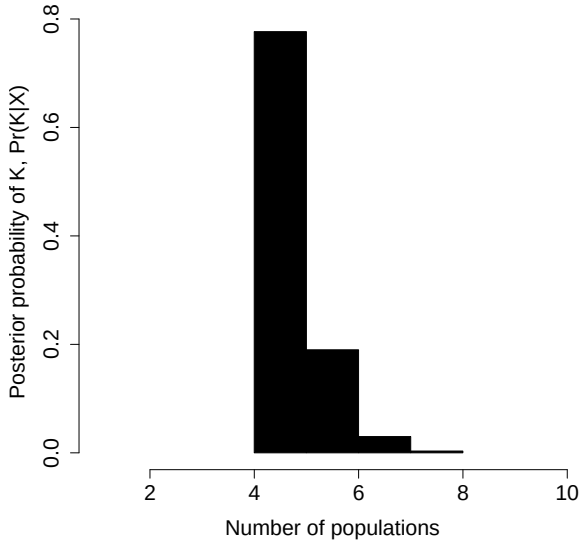


Figure 4.5: Posterior distribution for number of populations, $Pr(K|X)$ for the HGDP data.

Human Genome Diversity Project:

The Human Genome Diversity Project dataset we analyze consists of 1048 individuals from 53 world populations genotyped at 783 microsatellite loci. Along with genotype information, the individuals are also labeled with the geographical divisions to which they belong. Using *Structure*, Rosenberg et al. [2002] have previously analyzed the genotype data and found that the population clusters correspond to major geographical divisions of the world. We used *StructHDP* to reanalyze this data (without making use of the geographical information). The sampler was run for 20,000 iterations with the first 10,000 iterations discarded as burn-in. Samples were taken every 25 iterations to thin the Markov chain.

To determine the optimal number of ancestral populations, we examined the posterior distribution of the number of populations (K). Figure 4.5 shows the posterior distribution. We find the posterior distribution has a single mode at $K = 4$ and non-zero probability mass for values of K up to 8. For further analyses, we use the maximum-likelihood sample from the MCMC sampling, which has 4 ancestral population components.

The contributions of the four ancestral populations to an individual's genome can be repre-

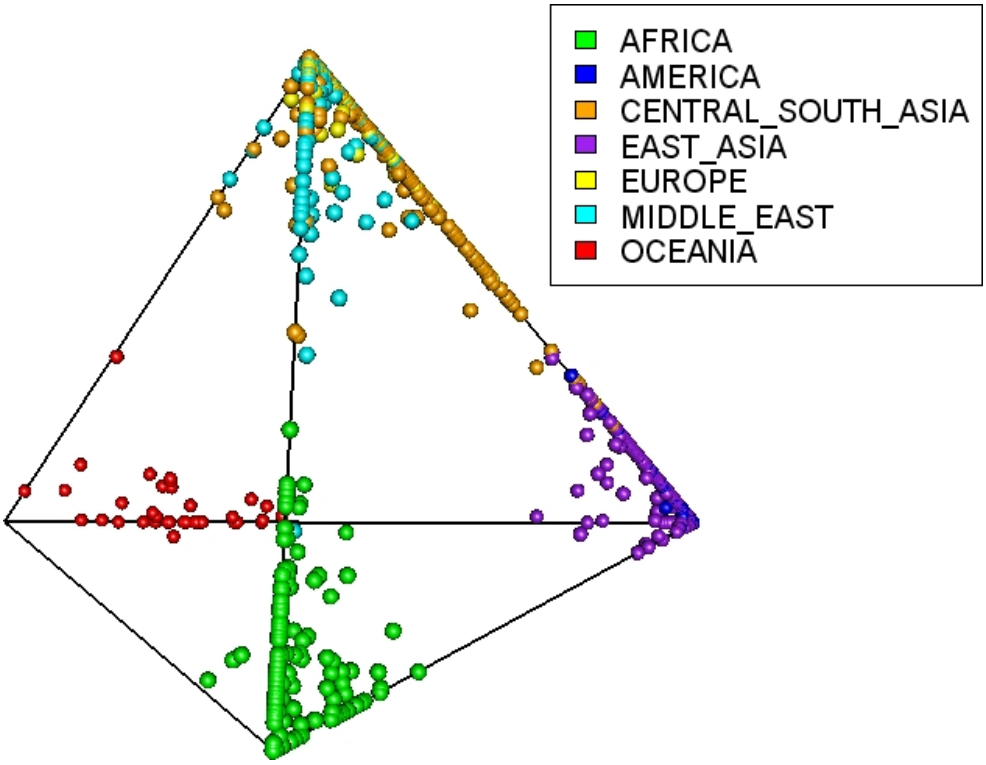


Figure 4.6: The ancestry proportions for the 1048 individuals from the Human Genome Diversity Project plotted in 3-dimensional space. Each individual is represented by a small sphere and the color of the sphere depends on the continental division the individual belongs to. Different colors correspond to different continental divisions. The geographical divisions are indicated by the labels on top of the graph.

sented using a 4-dimensional vector whose components sum to 1. All these vectors (referred to as ancestry proportions) lie within a tetrahedron in 3-dimensional space. Each of the four vertices of the tetrahedron represents an ancestral population. To visualize the clustering, we plotted the ancestry proportions for the 1048 individuals in 3 dimensions along with the tetrahedron in which the vectors lie. In this representation, the distance of a vector from the vertices of the tetrahedron indicates the amount of admixture present in an individual's genome. The further away from a vertex the vector is (and the closer it is to the center of the tetrahedron), the more the admixture present in the individual's genome.

Figure 4.6 shows the resulting plot for the 1048 individuals in the HGDP dataset. In the plot, each individual is represented by a small sphere. For ease of interpretation, the individual spheres are colored based on the geographical division they belong to. In the populations we examine,

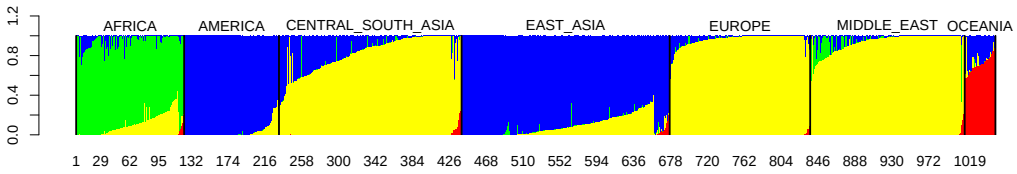


Figure 4.7: The ancestry proportions for the 1048 individuals from the Human Genome Diversity Project inferred by *StructHDP*. Each thin line denotes the ancestry proportions for a single individual. Different colors correspond to different ancestral populations. Dark black lines separate individuals from different major geographical divisions. The geographical divisions are indicated the labels at the top of the graph.

the divisions are Africa, the Americas, Central and South Asia, East Asia, Europe, Middle East and Oceania. These are represented by seven different colors. From the figure, we can see that individuals from a single continent cluster together in the same region of the tetrahedron. Some individual genomes are derived from a single ancestral population and lie at the vertices of the tetrahedron. Some other individuals, particularly those belonging to the Middle Eastern, Central Asian and South Asian populations, show a lot of admixture.

To analyze these results further, we plotted the ancestry proportions of the 1048 individuals as a bar graph, where every individual is represented by a thin bar with 4 components which sum to 1. Figure 4.7 shows the resulting bar graph.We can see that the clusters obtained correspond to the major geographical divisions of the world and the ancestral populations can be roughly described as *ancestral African* (denoted by green color), *ancestral American-East Asian* (blue), *ancestral European* (yellow) and *ancestral Oceanian* (red). From the ancestry proportions, we can see that the modern East Asian populations and American populations are similar, with the modern East Asian populations having a larger contribution from the ancestral population corresponding to Europe. Modern Asian populations also show some Oceanic ancestry (from the ancestral population denoted by red color). Modern Central and South Asian populations show an admixture of European and East Asian ancestral populations. The Middle Eastern populations show contributions from the ancestral African population and the ancestral European population.

Modern Oceanic populations are an admixture of an ancestral Oceanic population with an ancestral East Asian population. All of these observations are in agreement with previous analyses of the data by Rosenberg et al. [2002] and other studies of regional populations. We should note that the clusters inferred by *StructHDP* are not identical to the ones observed by Rosenberg et al. [2002] for $K = 4$, who observe that East Asia separates out into a separate cluster for $K = 4$ while Oceania separates from the rest of the data only for values of K larger than 4.

To analyze the similarity and differences within and between continental divisions, we computed the mean ancestry proportions for the 7 continental divisions by averaging the ancestry proportions for all individuals belonging to each continental division. We then constructed a distance matrix by computing the euclidean distance between the 4-dimensional vectors representing each continental division. Figure 4.8 shows the resulting distance matrix. From the figure, we can see that the distance matrix has a block structure. Modern American and East Asian populations are similar to each other and show little separation. We also see that modern European, Central-South Asian and Middle Eastern populations are close to each other. Within these 3 divisions, we see that Europeans and Middle Eastern populations group together while the Central-South Asians are further apart.

We hypothesized that if the inferred ancestry proportions capture the genetic variation between and across populations, then the pairwise Euclidean distance computed earlier should be correlated with genetic distance. To test this hypothesis, we computed the pairwise F_{st} distance between the 7 continental divisions of the data. To test for correlation between the pairwise Euclidean distance matrix and the pairwise F_{st} distance matrix, we used a Mantel test [Mantel, 1967]. A Mantel test tests the alternate hypothesis of correlation between two matrices against the null hypothesis of no correlation by permuting the rows and columns of one of the matrices and observing the distribution of the correlation statistic. The Mantel test on the Euclidean and F_{st} distance matrices shows that the correlation between the two distance matrices is 0.57 (P-value = 0.0025 with 10,000 replicates). The distribution of the observed and simulated Mantel

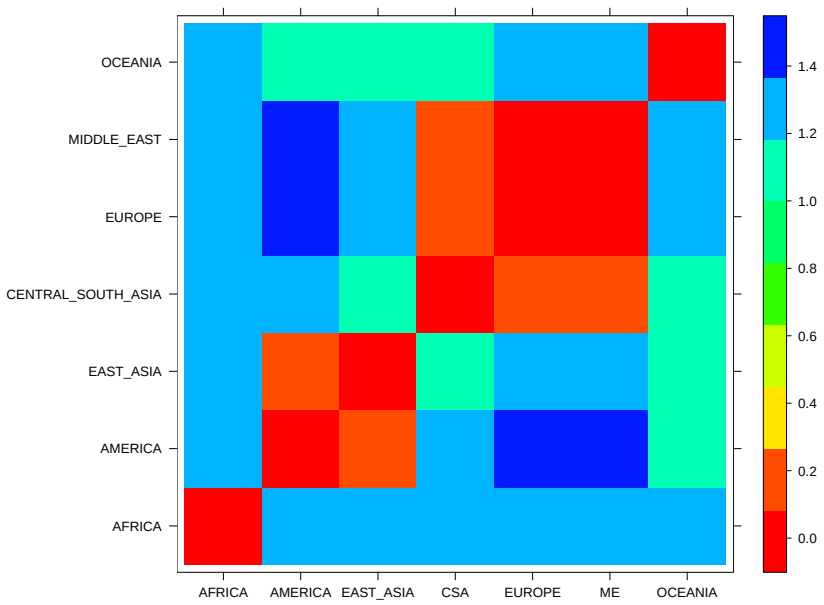


Figure 4.8: A matrix representing the distances between the mean ancestry proportions of the 7 major continental divisions of the HGDP. Red color indicates less distance while blue color indicates more distance.

correlation statistic is shown in Figure 4.9. Thus, we can see that the Euclidean distance and F_{st} distance are strongly positively correlated, which supports the inferred population structure.

To compare our results on the HGDP data with other methods, we analyzed the data using *Structurama*. However, due to computational reasons, we were unable to run *Structurama* on the full data at optimal settings. Therefore we analyzed a subset of the data that included only 100 loci per individual. We found that the posterior distribution of K inferred by *Structurama* has non-zero mass only at $K = 5$. Figure 4.10 shows the inferred ancestry proportions based on the mean partition from *Structurama*. We can see that *Structurama* also clusters the European, Middle Eastern and Central South Asian populations into a single cluster. However, since it does not allow partial membership, the individuals in different clusters have zero similarity. It is therefore unable to model the partial similarity between populations from different geographical divisions, for example, the Central Asian populations and European populations.

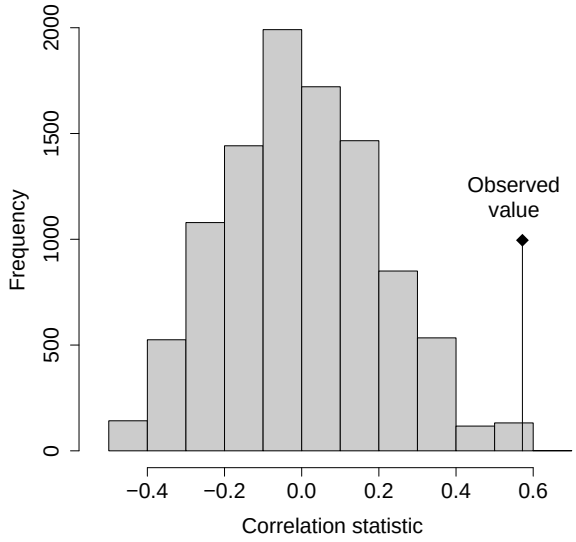


Figure 4.9: The distribution of the Mantel correlation statistic for the pairwise Euclidean distance matrix and the pairwise F_{st} distance matrix. The stem indicates the observed value of the statistic. The result is significant, with the associated P-value=0.0025

4.6 Discussion

We have presented *StructHDP*, a method for automatically inferring the number of population clusters present in a group of individuals while accounting for admixture between populations. At the same time, it also infers individual ancestry estimates under a *Structure*-like model. We demonstrated the effectiveness of our method on data simulated from an island model. We also analyzed the Taita thrush dataset and demonstrated that *StructHDP* chooses the number of clusters that best explain the data. Our analysis of the HGDP dataset shows that our method is able to cluster populations even when the individuals in the dataset are admixed. The ancestry proportions inferred for populations can be used to compute a distance measure between populations. We found that the Euclidean distance between populations has a strong positive correlation with the F_{st} distance between populations. The ancestry proportions therefore provide a useful low-dimensional representation of populations.

Our method uses a Hierarchical Dirichlet process to model the admixture of an unknown number of ancestral populations present in individual genomes in a given dataset. The HDP

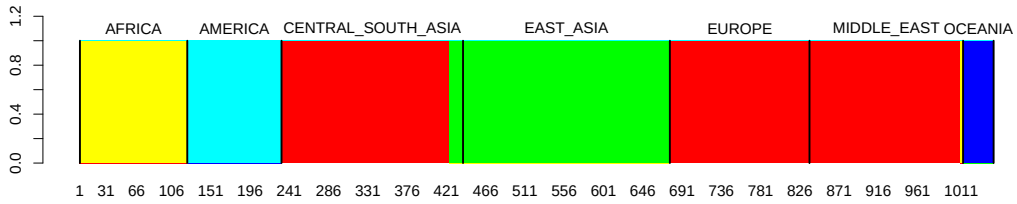


Figure 4.10: The ancestry proportions for the 1048 individuals from the Human Genome Diversity Project inferred by *Structurama*. Each thin line denotes the ancestry proportions for a single individual. Different colors correspond to different ancestral populations. Dark black lines separate individuals from different major geographical divisions. The geographical divisions are indicated the labels on top of the graph.

framework allows us to impose a Bayesian prior on the number of populations. We use an MCMC sampling algorithm, Gibbs sampling, to estimate the model parameters. The number of ancestral populations that best explain the data is one of the parameters of our model. The collapsed Gibbs sampler we implemented according to Teh et al. [2005] marginalizes the uncertainty in the population allele frequencies, thus eliminating a possible source of error in the inference. Our experiments suggest that the HDP is not sensitive to the priors on the parameters α_0 and γ since we sample them in the algorithm. The results are more sensitive to the choice of λ for the base distributions. A large value of λ tends to produce populations with uniform (high-entropy) allele frequency distributions while a small value of λ produces populations with allele frequency distributions highly skewed in favor of very few alleles (low-entropy).

The model as described here can handle both SNP and microsatellite markers. However, one of the limitations of our method is the computational time required for the Gibbs sampling. This means that while our method can handle datasets of a few thousand markers and individuals, it cannot be efficiently used on large datasets of hundreds of thousands of markers. However, as our simulations show, even with few loci, the method performs well at recovering the number of populations required to explain the data best. Teh et al. [2008] have described a way of implementing collapsed variational inference for HDPs. Applying the variational inference algorithm to *StructHDP* would improve its speed significantly.

In this work, we have shown how the basic admixture model can be extended to allow automatic inference of the number of populations. Just as extensions to the *Structure* model that account for recombination [Falush et al., 2003] and mutation [Shringarpure and Xing, 2009] have been developed, we can also extend *StructHDP* to model other evolutionary processes.

Genetic datasets are often accompanied by geographical information about the genotyped individuals. In some cases, there is a single geographical label associated with each individual, while in others, there are labels at different resolutions (for example, region, nation, continent). It has been shown that geographical distance correlates well with genetic distance between populations [Cavalli-Sforza et al., 1994, Novembre et al., 2008, Ramachandran et al., 2005]. Therefore the amount of sharing of ancestral population components between modern population groups is likely to depend on their geographical distance.

In its current form, *StructHDP* does not make use of geographical information in the inference process. Teh et al. [2005] describe how an HDP can be extended to include multiple levels of hierarchy and be generalized to a tree-like hierarchy. Use of the hierarchical geographical labels could allow us to impose a tree structure on the dataset that respects the geographical labels and enforces a level of population-sharing among individuals that is consistent with their geographical labels and distances.

Chapter 5

Effect of sample selection bias on population structure

Many public genotyping projects have made a large number of datasets available for population genetics studies. However, practical constraints dictate that of a geographical/ethnic population, only a small number of individuals are genotyped. The resulting data are therefore a sample from the entire population. If the distribution of sample sizes is not representative of the populations being sampled, the accuracy of population stratification analyses of the data could be affected.

We attempt to understand the effect of biased sampling on the accuracy of population structure analysis and individual ancestry recovery. We develop a mathematical framework to account for sample selection bias in models of population structure. We examined two commonly used methods for the analysis of such datasets - *Admixture* and *Eigensoft*. We found that the accuracy of population structure recovery by these methods is affected to a large extent by the sample used for analysis and how representative it is of the underlying populations. Using simulated and real data from the Human Genome Diversity Project, we show that sample selection bias can affect the results of population structure analyses. This is the first attempt at modeling sample selection bias in unsupervised clustering settings.

We propose a correction for sample selection bias using auxiliary information about the sam-

ple. We demonstrate that such a correction is effective in practice using simulated and real data.

5.1 Introduction

A large number of genetic datasets such as the HAPMAP [Gibbs, 2003], Human Genome Diversity Project (HGDP) [Cavalli-Sforza, 2005] and others are available for the study of population structure. Many datasets which sample a number of individuals from a specific region have also been analyzed to look for evidence of population stratification. These datasets contain individuals from many geographically and ethnically diverse populations. Due to practical constraints, only a small number of individuals from each population are genotyped and the resulting data form a sample from the entire population. This often means that the sample selected for analysis is a biased sample from the underlying populations. This problem is also encountered when multiple datasets are combined to detect population structure with better resolution.

We hypothesize that if the distribution of sample sizes is not representative of the populations being sampled, the accuracy of population stratification analyses of the data could be affected. This is because a fundamental assumption of many statistical learning algorithms is that the sample available for analysis is representative of the entire population distribution. While most algorithms are robust to minor violations of this assumption, sampling bias in the case of genetic datasets may be too large for algorithms to accurately recover stratification.

Our results on simulated data show that accuracy of population stratification and recovery of individual ancestry are affected to a large extent by the sampling bias in the data collection process. Both likelihood-based methods and eigenanalysis show sensitivity to the effects of sampling bias. We show that sample selection bias can affect population structure analysis of the HGDP data, leading to potentially incorrect interpretations of evolutionary history. We also propose a mathematical framework to model sample selection bias, and a correction that can reduce its effects. We show how such a correction can be implemented and its effectiveness in practice.

5.2 Related work

In this section, we briefly examine the factors that affect the accuracy of population stratification methods. We also examine related work on addressing the problem of sample selection bias in various contexts and demonstrate that sample selection bias may exist in genetic datasets.

5.2.1 Factors affecting accuracy of stratification

A number of factors are known to affect the accuracy of population stratification and individual ancestry recovery. In one of the early studies on model-based methods for population stratification, Pritchard et al. [2000a] showed that the number of loci available for analysis had a significant effect on the recovery of individual ancestry using *Structure*. Kaeuffer et al. [2007] studied the effect of linkage disequilibrium on recovery of population structure using simulated data. McVean [2009] suggested an interpretation of the eigenanalysis method that is the basis of the *Eigensoft* method in terms of the coalescence times of individuals. They also explored many scenarios in which eigenanalysis performs well or badly. In the following subsection, we discuss the problem of sample selection bias and some related work on the effect of biased sampling on population stratification accuracy.

5.2.2 Sample selection bias

A common assumption of many statistical algorithms is that the available sample is representative of the underlying population. In reality, however, this assumption may not always be correct. Sample selection bias is any systematic difference between the sample and the population. It affects the internal validity of an analysis by leading to inaccurate estimation of relationships between variables. It can also affect the external validity of an analysis since the results from a biased sample may not generalize to the population.

The problem of sample selection bias was first widely studied in econometrics, where it ap-

peared as a bias among survey responders. Heckman [1979] provided a method of addressing this problem in linear regression models by estimating the probability of an individual being included in the sample. Sample selection bias has also been addressed in statistics and machine learning literature by attempts to understand its effect on various classifiers and how estimation and prediction can be made correctly in the presence of sampling bias [Cortes et al., 2008, Davidson and Zadrozny, 2005, Vella, 1998, Zadrozny, 2004]. Zadrozny [2004] discusses the properties of various learning algorithms and the effect of sample selection bias on their accuracy. It also outlines a possible way of correcting for sample selection bias provided we know the nature and structure of the bias. Sample selection bias is also studied in ecology when trying to model species distributions using presence-only data [Phillips et al., 2009]. An alternative view of sample selection bias is provided by the statistics literature examining the problem of incorporating sampling weights in models. Bertolet [2008] examines the problem of incorporating sampling weights in mixed-membership models similar to the models we examine here.

To demonstrate that existing genetic datasets show evidence of sample selection bias, we use data from the HGDP. It is important to note that in the absence of knowledge of the underlying distribution over genotypes (which are very high-dimensional and therefore have very complex distributions) or the underlying true ancestries (presumed to be low dimensional and therefore easier to characterize), an exact quantification of sample selection bias is impossible in real datasets. However, since it has been shown that geographic distance correlates well with genetic distance [Ramachandran et al., 2005], we will use geographic labels as proxy for true ancestry in this demonstration. The HGDP includes continent, nation and geographic region labels for every individuals. We choose to use nation labels as the proxy for true ancestry here since the data for the true population of nations is readily available. Figure 5.1 shows the plot of the population of a country against the number of individuals from that country genotyped in the HGDP. An unbiased sample according to population size should allow a good linear fit to the graph. However, we observe that the linear fit is not good ($R^2 = 0.22$). This suggests that the

HGDP sample is biased. This bias could be due to constraints on the sampling, or by design (to obtain more data about certain groups which are of more evolutionary interest).

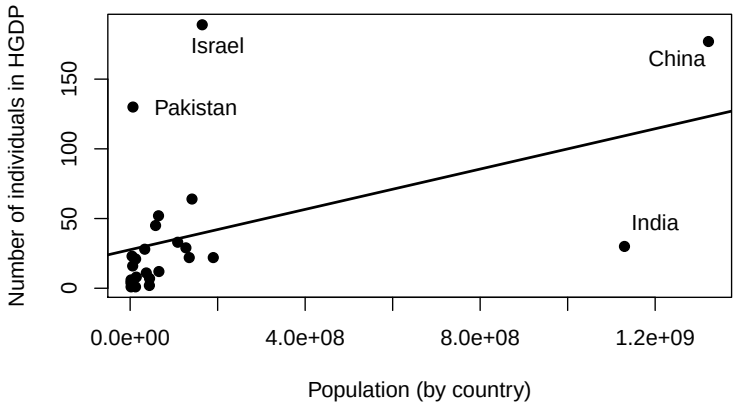


Figure 5.1: Plot of population vs number of individuals in the HGDP, by country. A line fit to the graph gives $r^2 = 0.22$. Four outliers, which are overrepresented or underrepresented in the sample compared to the expected number by the linear fit, are labeled by their country names.

In population genetics, sample selection bias could be a serious problem since the estimates of ancestry obtained from stratification analyses are often used to make inferences about the evolutionary history of populations. The inferred individual ancestries are also used as input in correcting for stratification in association studies (for instance, in *Admixmap* [Hoggart et al., 2003]). Pritchard et al. [2000a] suggest that detecting stratification is difficult unless a significant number of unmixed individuals from each ancestral (or pseudo-ancestral) population is present in the sample. This observation was verified by Tang et al. [2005] through experiments on a small number of simulated datasets. To our knowledge, there has been no systematic study of the effect of sample selection bias on the accuracy of population structure recovery and individual ancestry recovery.

We propose to study the effect of sample selection bias on the accuracy of population stratification and individual ancestry recovery using both a model-based approach (*Admixture*) and an eigenanalysis-based approach (*Eigensoft*). Since the analysis of McVean [2009] provides guidelines on the effect of sampling bias on stratification accuracy using eigenanalysis, we will focus

our attention mainly on probabilistic models such as *Admixture*.

5.3 A mathematical framework for sample selection bias

We consider the problem of studying genotype data using a probabilistic model. For probabilistic modeling, we would ordinarily assume that we have genotypes (g) drawn independently from a distribution D (with domain $\mathcal{G}$) over the feature space G. We assume that our points (g, u, s) are drawn independently from a distribution D over $\mathcal{G} \times \mathcal{U} \times \mathcal{S}$, where G is the space of genotypes, U are some auxiliary features of the data that are not of direct interest for modeling and S is a binary space. The variable s controls the selection of points (1 means the point is selected and is observed in our sample, 0 means the point is not selected). Our observed sample contains only points that have $s = 1$. We will refer to this as the selected sample and refer to its distribution as D'.

We consider the setting where s is independent of g given u, that is $P(s|g, u) = P(s|u)$. This setting, where the selection is controlled by features different from the genotype we want to model, arises frequently in real applications. In population genetics, whether an individual is included in a genotyping study often depends on factors such as geographical location.

5.3.1 Sample selection bias correction

It is evident that if g is independent of u in the previous setting, then sample selection bias has no effect and the probability of x in the selected sample is the same as probability of g under D (asymptotically). If g and u are not independent, then we can write using Bayes rule:

$$P(g, u) = \frac{P(s = 1)P(g, u|s = 1)}{P(s = 1|g, u)} \tag{5.1}$$

$$= \frac{P(s = 1)P(g, u|s = 1)}{P(s = 1|u)} \tag{5.2}$$

which can be rewritten as

$$P_D(g,u) = \frac{P(s=1)P_{D'}(g,u)}{P(s=1|u)} \tag{5.3}$$

where D' represents the distribution of the selected sample. Since the term $P(s=1)$ is constant with respect to (g,u), we can say that

$$P_D(g,u) = \frac{c \times P_{D'}(g,u)}{P(s=1|u)} \tag{5.4}$$

where c is a constant that need not be evaluated for tasks such as learning model parameters. Therefore, to model $P_D(g,u)$ accurately (upto a multiplicative constant), we can follow the procedure below:

1. Compute $P_{D'}(g,u)$ using a model learned on the selected sample.
2. Apply a correction using the term $P(s=1|u)$. This can be done in two ways:
   (a) If we know the selection procedure, we know $P(s=1|u)$ and can directly use it.
   (b) If we don't know the selection procedure, but we have access to large number of points for which we know (u,s), but not g, we can estimate $P(s=1|u)$. In the population genetics example, this would correspond to knowing the income or geographical region of an individual and whether or not they could have been included in the study (genotyping individuals to find g for a large sample would be expensive).

However, this analysis, which can accurately correct for sample selection bias in the described setting, requires a model of both g and u. In many applications, we are interested in only modeling g and not u. For instance, while there is interest in modeling the distribution of genotypes, distributions of income or geography are not of interest in genetics. Therefore, we consider a similar analysis in the case where we only model $P(g)$ and attempt to derive a correction for sample selection bias.

5.3.2 Approximate correction

We consider the case when we only want to model $P(g)$. Proceeding in a similar way as before, we can write:

$$P(g) = \frac{P(s=1)P(g|s=1)}{P(s=1|g)} \tag{5.5}$$

which can be restated as:

$$P_D(g) = \frac{c \times P_{D'}(g)}{P(s=1|g)} \tag{5.6}$$

A correction for sample selection bias could therefore be found if we could estimate $P(s = 1|g)$. However, g is typically high-dimensional — in genetics applications, g may have dimensions from 1,000-1,000,000. $P(s = 1|g)$ is therefore hard to estimate from the small selected sample. We propose that since g and u are dependent and u typically has much lower dimensionality than g, we can approximate $P(s = 1|g)$ by $P(s = 1|u)$. We can therefore write the correction for sample selection bias as

$$P_D(g) \approx \frac{c \times P_{D'}(g)}{P(s=1|u)} \tag{5.7}$$

with the quality of the approximation varying as a function of the dependence between g and u. In practice, we find that the approximate correction method is adequate for most applications, since probabilistic models are often robust to some differences between the true distribution of the data and the distribution of the selected sample.

It is important to note that even if the selection is determined in reality by the u variables only, the correction proposed in Equation 5.7 is only an approximate correction. The exact correction would require computing the term $P(s = 1|g)$ which can be written as $\sum_u P(s = 1|u)P(u|g)$. The second term is a distribution conditioned on g and is hard to specify due to the high dimensionality of g.

5.3.3 Implementing correction in learning

While applying the proposed correction for accurate probability modeling only requires an extra multiplication step, implementing the correction in learning models consistent with the true distribution is more complex. This problem is well-studied as cost-sensitive learning. A discussion of the ways in which the correction can be applied to classifiers can be found in Zadrozny et al. [2003]. In this work, we will use sampling with replacement to implement the correction. To perform sampling with replacement, we sample points in the selected sample (with replacement) with probability proportional to their correction factor ($1/p(s = 1|u)$). If the selected sample contains N points, the probability of inclusion for the i^{th} point is given as $\frac{1/p(s=1|u_i)}{\sum_{j=1}^{N} 1/p(s=1|u_j)}$. Since we sample with replacement, our corrected sample can include non-unique points from the selected sample.

5.4 Methods

We will demonstrate the effects of sample selection bias on the accuracy of ancestry recovery using experiments on simulated and real data. We will also show how the approximate correction for sample selection bias is effective in practice.

5.4.1 Simulation experiments

To examine the recovery of individual ancestry, we simulated data depicting the scenario shown in Figure 5.2. In this scenario, a population P_0 of size N at mutation-drift equilibrium splits into two isolated subpopulations P_1 and P_2, each of size N. For a number of generations (that can be varied as a parameter), the two populations have no gene flow between them. Finally, the two populations are pooled and random mating takes place for G generations in the combined population. This allows us to record the true ancestry of every individual in the resulting sample P.

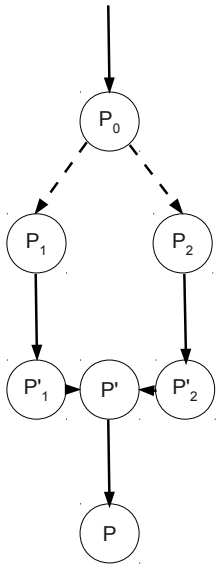


Figure 5.2: Simulation scenario for data generation.

For our experiments, we set the size of the original population to 1000 (N=1000). We set the mutation rate to 10^{-9} per site per generation and the recombination rate to 10^{-8} per site per generation and generated a region with 50,000 SNPs. The generated ancestral populations had high F_{ST} differentiation (mean F_{ST}=0.44). The F_{ST} and number of loci were chosen to be high to avoid the confounding effects of these two factors on the accuracy of ancestry inference. The pooled populations were randomly mated for a single generation (G=1). For statistical significance, we generated 30 datasets using the simulation settings. We used the coalescent software *Genome* [Liang et al., 2007] to generate the two ancestral populations.

For each diploid individual in the resulting population, two parents were randomly chosen from the pool. Therefore, in expectation, 25% of the resulting individuals will have both parents from population 1, 25% will have both parents from population 2 and 50% will have a parent each from both populations. We generated 1000 individuals from the random mating. We use the 2-dimensional ancestry vector $(\theta_i, 1-\theta_i)$ to represent the contributions from the two ancestral populations to the genome of the i^{th} individual. The generated population contained the following three groups:

1. 250 unmixed inds. with ancestry from the first ancestral population, $(\theta_i, 1-\theta_i) = (1, 0)$
2. 250 unmixed inds. with ancestry from the second ancestral population, $(\theta_i, 1-\theta_i) = (0, 1)$
3. 500 admixed inds. with ancestry from both populations. While the specific proportions of ancestry within this group vary, we have that $E[(\theta_i, 1-\theta_i)] = (0.5, 0.5)$

To study the sampling bias, we subsample individuals from the dataset to generate a dataset of size x+y+z where x is the number of individuals with both parents from population 1, z is the number of individuals with both parents from population 2, and y is the number of individuals with one parent from each population. By varying x, y, and z, we can generate smaller datasets with different kinds of bias and deviations from the original dataset. We choose x and z from {5,10,25,100,250} and y from {5,10,25,100,250,500}. Thus the smallest possible dataset is {5,5,5} (15 individuals) and the larget possible dataset is the same as the original dataset {250,500,250} (1000 individuals). We will use S_{xyz} to refer to the dataset $\{x, y, z\}$.

In this case, g represents the genotypes of the individuals (which are 50,000-dimensional), u represents the group memberships of each individual according to their ancestry ($u \in \{1, 2, 3\}$, with the 3 groups as defined earlier). By design, s depends only on u, and we can write the probability distribution $P(s = 1|u)$ as:

$$P(s = 1|u = 1) = x/250 \tag{5.8}$$

$$P(s = 1|u = 2) = z/250 \tag{5.9}$$

$$P(s = 1|u = 3) = y/500 \tag{5.10}$$

5.4.2 Evaluation measure

A fair evaluation of the results for both *Admixture* and *Eigensoft* is difficult to achieve because the individual ancestries produced by *Admixture* and *Eigensoft* are different in nature. With K ancestral population, an individual ancestry vector produced by *Admixture* has the form $\{q_1, \cdots, q_K\}$ such that $\sum_{k=1}^{K} q_k = 1$. Thus it has only $K - 1$ independent components. With K eigenvectors,

an individual ancestry vector produced by *Eigensoft* is the projection of the genotype of the individual on the K eigenvectors. It has no restrictions, unlike the *Admixture* ancestry vectors, and has K independent components. The ancestry vectors that we store as the true ancestry when generating the simulation data have the same form as those produced by *Admixture*.

To reduce the effects of the different natures of inferred ancestries on our evaluation, we devised an evaluation measure that depends only on the distances induced by the ancestry. Suppose two individuals i and j have ancestry vectors $q^i = \{q^i_1, \cdots, q^i_K\}$ and $q^j = \{q^j_1, \cdots, q^j_K\}$ respectively in K dimensions (ancestral populations in *Admixture* or eigenvector projections in *Eigensoft*). The euclidean distance between their ancestries is given by $\|q^i - q^j\|_2 = \sqrt{\sum_{k=1}^{K}(q^i - q^j)^2}$. Therefore, given a set of ancestry vectors for a dataset S, we can compute the distance matrix induced by the ancestry vectors computed using a particular method. We denote the distance matrix induced on dataset S by the *Admixture* ancestry vectors as $D^S_{\textit{Admixture}}$ and the distance matrix due to the *Eigensoft* ancestry vectors as $D^S_{\textit{Eigensoft}}$. If D^S_{true} represents the distance matrix of the true ancestry vectors, measuring the magnitude of the correlation between the distance matrices gives us a measure of the accuracy of recovery of individual ancestry that should be agnostic of the method used to infer ancestry. To evaluate the effect of biased sampling on the accuracy of *Admixture*, we will examine the effect of varying x, y and z on $|\text{Correlation}(D^{S_{xyz}}_{true}, D^{S_{xyz}}_{\textit{Admixture}})|$. For *Eigensoft*, we shall do the same using $D^{S_{xyz}}_{\textit{Eigensoft}}$. For statistical soundness, we report the mean of the absolute value of the correlation over 30 datasets simulated using the same parameters.

An alternative evaluation metric that is more intuitive for likelihood-based methods is discussed in the Appendix (Section 5.6). However, this metric does not generalize to eigenanalysis methods and therefore we do not use it in our analyses.

5.5 Results

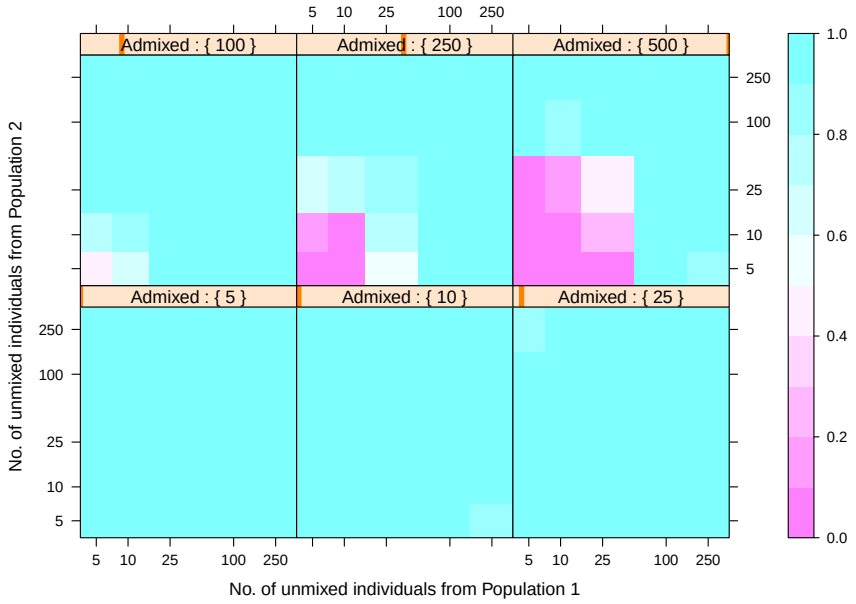
We examined the effect of biased sampling of individuals by constructing subsets of the whole dataset and measuring the correlation between the distances induced by the true and inferred

ancestry. Figure 5.3(a) shows the results of this analysis with the *Admixture* software with 6 sub-plots. Within each sub-plot, the number of admixed individuals in the sample remains constant and the number of unmixed individuals from the two ancestral populations is varied. Figure 5.3(b) shows the results of an identical analysis of the same dataset with the *Eigensoft* software.

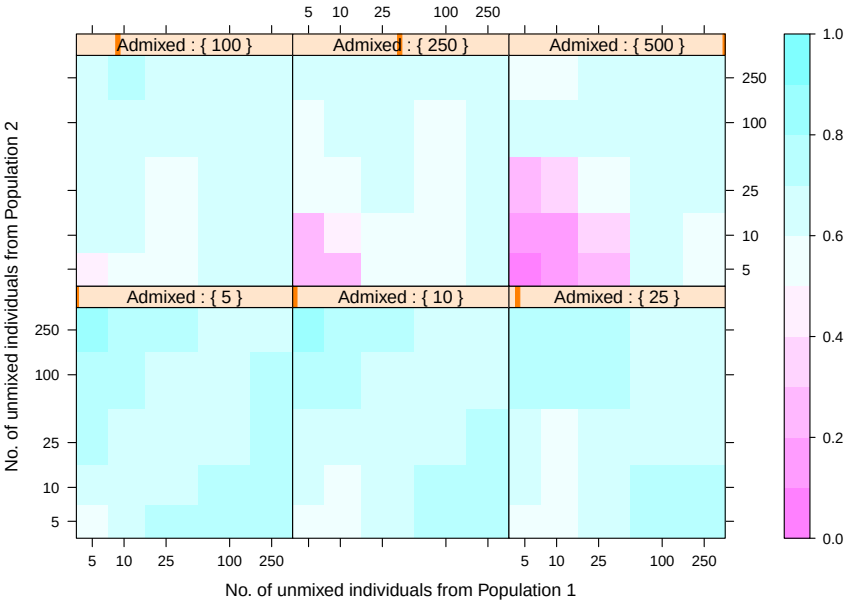
We observed some common trends in the results of both analyses. Overall, both methods recovered individual ancestry well, with the average correlation being 0.92 for *Admixture* and 0.62 for *Eigensoft*. Previous work has shown that the accuracy of individual ancestry recovery is a function of the F_{ST} differentiation between the ancestral populations. We shall therefore note that the results we obtain may vary for datasets with different F_{ST} values between the ancestral populations and quantifying this effect will require more study. The simulation setting we described earlier generates ancestral populations that are easily separable even when we have access to little data. This can be observed in Figures 5.3(a) and 5.3(b).

However, in the scenario where we have few unmixed individuals from both ancestral populations, Figures 5.3(a), 5.3(b) show that the accuracy of individal ancestry recovery drops significantly. This effect is noticeable in the sub-plots with 250 and 500 admixed individuals. In all sub-plots, the results show no noticeable drop in accuracy when we have 100 or more unmixed individuals from at least one ancestral population. Pritchard et al. [2000a], Tang et al. [2005] have previously noted that a significant number of unmixed individuals from each ancestral population is required for accurate recovery of stratification. An initial examination of the results suggests that it may be sufficient to have a large number (around 50-100) of unmixed individuals from just one of the two ancestral populations to be able to correctly resolve stratification.

We note that this guideline, which is relevant when the number of admixed individuals is large, does not apply if the dataset contains few admixed individuals and few unmixed individuals. When there are few admixed individuals, both methods perform well (relative to their best performance) even with 5 unmixed individuals in the dataset. Mantel tests [Mantel, 1967]



(a) Using *Admixture* with K=2



(b) Using *Eigensoft* with top two eigenvalues

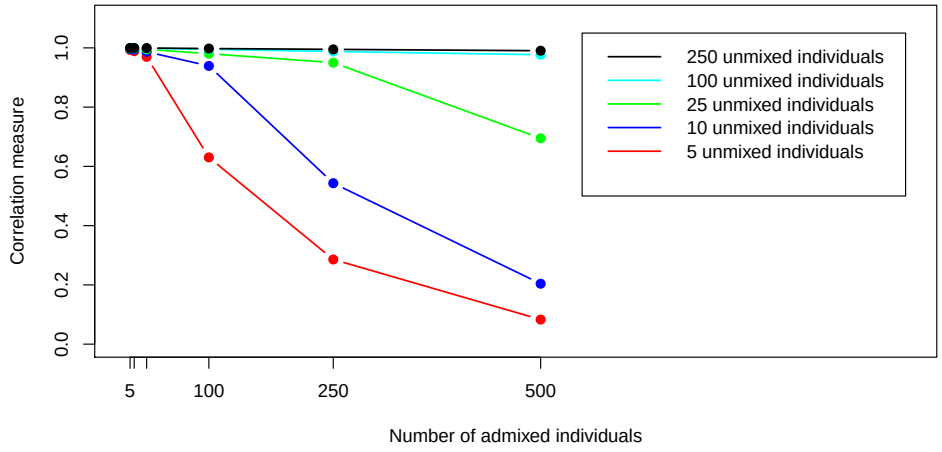
Figure 5.3: Correlation between the true individual ancestry and the individual ancestry inferred using (a) *Admixture* with K=2 and (b) *Eigensoft* with the top two eigenvalues. The different levelplots are drawn for different number of admixed individuals in the dataset. The X and Y axes of the plots are logarithmic in scale.

over the resulting distance matrices reveal that the high correlations obtained with few admixed individuals are statistically significant ($p < 10^{-3}$) in all cases.

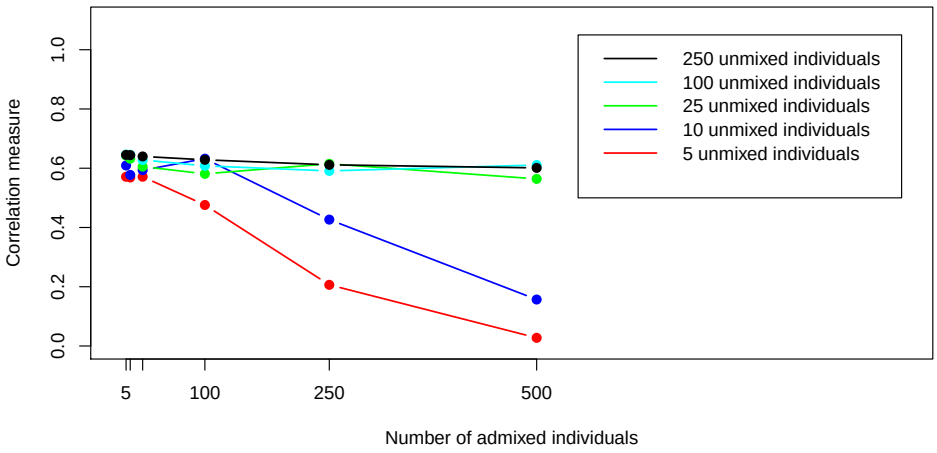
To examine the effect of the number of unmixed individuals on the ancestry recovery in more detail, we looked at the subset of the generated subsamples which had the same number of unmixed individuals from both ancestral populations, i.e., subsets of the form S_{xyx} where the number of unmixed individuals x$\in\{5,10,25,100,250\}$ and the number of admixed individuals y$\in\{5,10,25,100,250,500\}$. For each value of x (the number of unmixed individuals from each ancestral population present in the dataset), we observed the effect of varying y (the number of admixed individuals present in the dataset) on the ancestry recovery. Figure 5.4(a) shows the results for *Admixture* and Figure 5.4(b) shows the results for *Eigensoft*. When the number of unmixed individuals is large, the methods recover ancestry well and the number of admixed individuals have no effect on accuracy. However, when the number of unmixed individuals in the sample is small, adding more admixed individuals to the sample reduces the accuracy of the ancestry recovery for both *Admixture* and *Eigensoft*. In Figures 5.4(a), 5.4(b) we see a threshold effect due to the number of unmixed individuals when the number of unmixed individuals changes from 25 to 100.

The high accuracy of ancestry recovery when there are few admixed and few unmixed individuals suggests that previous intuition about the requirement of a large number of unmixed individuals for accurate ancestry recovery may be an incomplete explanation. The results in Figures 5.3(a) and 5.4(a), along with the likelihood model underlying *Admixture*, suggest that the effect on accuracy may depend on the ratio of the number of admixed individuals to unmixed individuals from each population in the sample. For notational convenience, we will refer to this ratio as $\tau_{sample} = y/x$.

To examine this hypothesis, we replot the data used for Figure 5.4(a) by examining the correlation measure as a function of the ratio of admixed individuals to unmixed individuals in the sample. Figures 5.6(a) and 5.6(b) show the results for this visualization for *Admixture* and



(a) Using *Admixture*



(b) Using *Eigensoft*

Figure 5.4: Effect of adding more admixed individuals to the dataset on the correlation measure of accuracy when using (a) *Admixture* with K=2 and (b) *Eigensoft* with the top two eigenvalues. The X axis is logarithmic in scale.

Eigensoft respectively. From the figure, we can see that the effect of sample selection bias can be understood using τ_{sample}. The accuracy of ancestry recovery is high while the value of τ_{sample} is less than 10 and drops as this ratio increases. This behavior is independent of the exact number of unmixed individuals in the dataset and can be observed for both *Admixture* and *Eigensoft*. An oversampling experiment using *Admixture* (Section 5.5.1 below) showed that even with 100 unmixed individuals from both ancestral populations, the correlation measure drops to 0.15 when τ_{sample} was increased to 25.

5.5.1 Oversampling experiment to demonstrate the effect of τ_{sample}

Our experiments show that the effect of sample selection bias on accuracy of population stratification depends on τ_{sample}, the ratio of admixed individuals to unmixed individuals (from each ancestral population) in the sample. To verify our hypothesis that this effect is independent of the exact number of unmixed individuals in the sample, we performed an oversampling experiment starting with subsets $S_{100,10,100}$. We oversampled the 10 admixed individuals from the sample while keeping the number of unmixed individuals fixed to obtain τ_{sample} values of 5,10,25,50, and 100. Figure 5.5 shows the results of the oversampling experiment reporting means over 30 datasets. From the figure we can see that the correlation measure of accuracy drops to 0.5 when $\tau_{sample} = 10$ and decreases to around 0.15 as τ_{sample} reaches 25 or higher values. This verifies our hypothesis that the effect of τ_{sample} on accuracy can be observed regardless of the exact number of unmixed individuals in the sample. The observed drop in accuracy is sharper than in Figure 5.6(a) due to the effects of oversampling.

As described in the simulation settings, the ratio of the number of admixed individuals to the number of unmixed individuals in the entire population, $\tau_{population}$, has expected value 2. In our experiments, we observe that individual ancestry can be recovered perfectly even when $\tau_{sample} > 2$ as long as $\tau_{sample} < 10$. A deficit of admixed individuals, indicated by $\tau_{sample} < 2$ has no adverse effect on the accuracy of ancestry recovery. The effects of sample selection

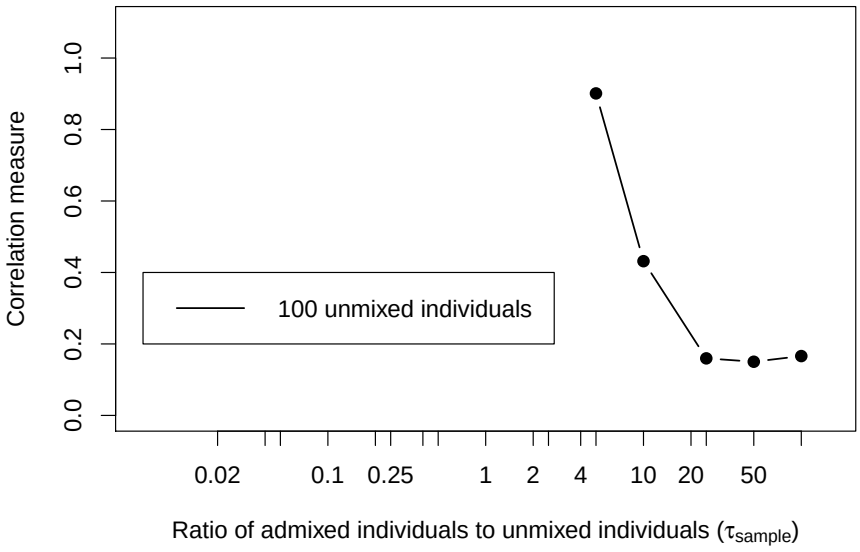


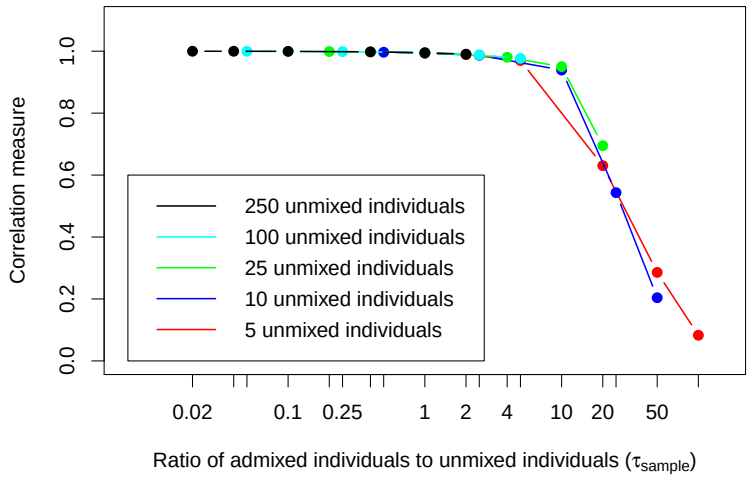
Figure 5.5: Effect of τ_{sample} with 100 unmixed individuals in an oversampling experiment. Admixed individuals are oversampled from 10 to obtain the desired value of τ_{sample}.

bias on the accuracy of ancestry recovery in a simple two-population admixture scenario using mixed-membership models can thus be explained in two scenarios: (i) when $\tau_{sample} < 10$, sample selection bias has no effect on the accuracy of individual ancestry recovery and (ii) when $\tau_{sample} > 10$, the accuracy of individual ancestry measured using the correlation measure decreases logarithmically with τ_{sample}.

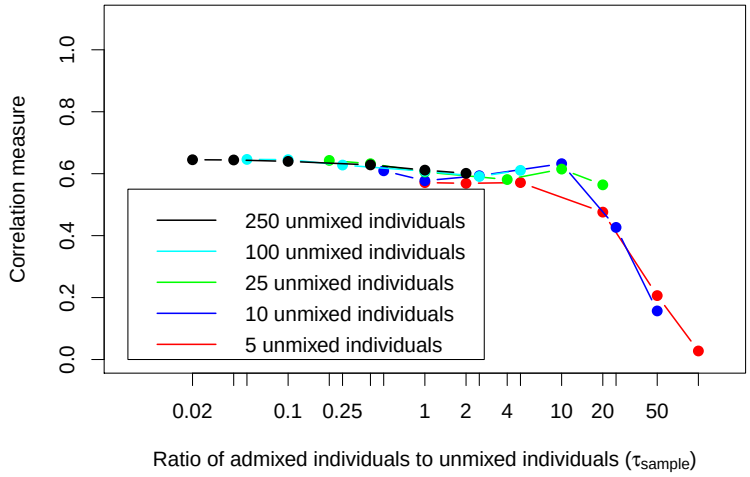
5.5.2 Comparing results of *Structure* and *Admixture*

To verify that the results we observed were a characteristic of the model underlying *Admixture* and not a result of the optimization method, we replicated our experiments using *Structure*. However, due to the high computational cost of *Structure*, we had to reduce the number of loci from 50,000 to 500 for all datasets.

Figure 5.7 shows the results of the *Structure* in the same format as the results for *Admixture*



(a) Using *Admixture*



(b) Using *Eigensoft*

Figure 5.6: Effect of the ratio of the number admixed individuals to the number unmixed individuals in the dataset (τ_{sample}) on the correlation measure of accuracy using (a) *Admixture* with K=2 and (b) *Eigensoft* with the top two eigenvalues. The X-axis is logarithmic in scale.

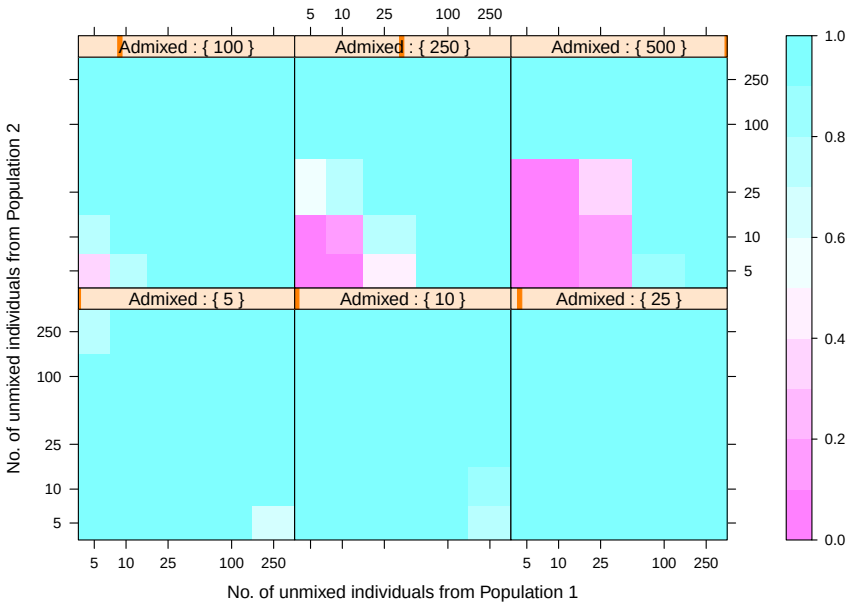


Figure 5.7: Correlation between the true individual ancestry and the individual ancestry inferred using *Structure* for K=2. The different levelplots are drawn for different number of admixed individuals in the dataset. The X and Y axes of the plots are logarithmic in scale.

in Figure 5.3(a). We can see that the two figures are similar, with the mean accuracy of *Structure* (0.91) slightly lower than that of *Admixture* (0.92). This suggests that the results are not an artifact of the different optimization methods chosen in the two methods.

Correction by resampling

From Equation 5.7, we can see that an approximate correction can be applied to the selected samples using the weights from Equation 5.10. We use the sampling with replacement method to implement a correction. On implementing such a correction, we found that the correlation measure of accuracy was larger than 0.99 in 99% of the corrected datasets. Since the corrected datasets only used the genomes of individuals present in the biased samples, we can infer that the loss in accuracy observed earlier was due to biased sampling.

5.5.3 Sample selection bias in the HGDP data

To demonstrate the effects of sample selection bias on a real dataset, we analyzed data from the HGDP. We used individuals from the HGDP for which the national labels were unambiguous and the national population data was readily available. We also ignored any SNPs which had missing data. The resulting dataset had 918 individuals from 24 countries genotyped at 2810 SNPs. We used the mixed-membership model of *Admixture* with $K = 5$ to produce low-dimensional ancestry representations for each individual since previous work suggests that $K = 5$ adequately demonstrates variations in human populations at the continental level. To analyze the results, we used the nationality labels for individuals to construct a mean ancestry representation for each of the 24 nations. A distance matrix was then constructed using these low-dimensional representations. Figure 5.8(a) shows the results of this analysis. From the figure, we can see that the national populations cluster by their continental locations, with much lower distances between two nations within a continent than two nations in different continents.

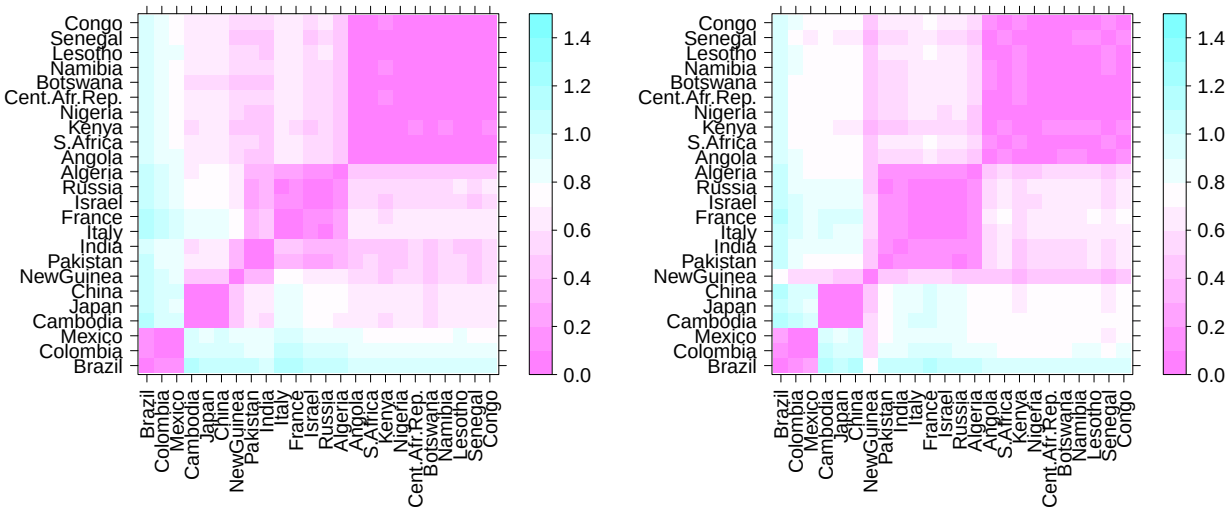
5.5.4 Correction for HGDP data

As we previously described, the number of individuals sampled from each country in the HGDP is not well-correlated with the population of that country. This bias can cause the results of ancestry inference on this data to not be representative of the underlying populations. In this case, we assume that the variable u, on which the selection procedure is based, is the country of origin for the individual. Let $n(u)$ denote the number of individuals from the country u included in the HGDP, $N(u)$ be the population of that country and N be the population of the world. Then, we can evaluate the correction probability for the HGDP dataset using:

$$p(s = 1|u) = \frac{n(u)}{N(u)} \tag{5.11}$$

We can apply this correction using sampling with replacement and re-analyze the HGDP data as described earlier with $K = 5$. Figure 5.8(b) shows the results of this analysis. We see that while

the results of the analysis are similar to the uncorrected results in Figure 5.8(a), there are some important differences between them. The uncorrected analysis suggests that the populations in the Indian subcontinent (India and Pakistan) are similar to the European populations as well as the East Asian populations (China, Japan and Cambodia) while the corrected analysis suggests that Indian and Pakistani populations are much closer to European populations than to the East Asian populations. Recent work on a large dataset from Indian populations supports the claim that populations in the Indian subcontinent are genetically more similar to European populations than to East Asian populations [Reich et al., 2009]. The analysis with the correction also separates the East Asian populations from the African populations more distinctly than the uncorrected analyses. While it is difficult to argue that the correction provides objectively better results, it is clear that due to sample selection bias, there are differences between the two samples and analyses that could lead to different interpretations of evolutionary history.



(a) Uncorrected

(b) Corrected

Figure 5.8: Analysis of distance between low-dimensional representations of national populations using *Admixture* for $K = 5$. (a) Original HGDP data, without correction (b) With correction for sample selection bias. The nations are sorted by their continental location.

5.6 Discussion

Our experiments suggest that sample selection bias can be a problem in accurate population stratification and recovering individual ancestry. In most stratification analyses, the recovered ancestry is used to make inferences about the evolutionary history of the underlying populations. It is also used in association studies to account for the effects of stratification. Therefore, it is essential to have accurate recovery of ancestry. Our simulations show that unlike observations from previous studies, the accuracy of individual ancestry recovery is not dependent only on the number of unmixed individuals present in the sample. We observe a threshold effect where the accuracy of ancestry inference is affected by sample selection bias depending on the ratio of admixed individuals to unmixed individuals in the sample. Our simulations showed that the accuracy of ancestry inference is affected when this ratio, τ_{sample}, is less than 10. However, more analysis is needed to determine whether this guideline is applicable in all scenarios and may differ depending on the F_{ST} differentiation between ancestral populations, number of loci available and other factors.

While our analyses used two specific methods (*Admixture* and *Eigensoft*), we claim that the effects we observed are a feature of the assumptions underlying both methods rather than the specific implementations. In the case of likelihood-based models, we demonstrated this by developing a probabilistic framework for sample selection bias. *Admixture* is a representative of the likelihood-based models that assume: (a) admixture between ancestral population and (b) that modern individual genomes are mixture of contributions from different ancestral populations. This is the model underlying *Structure* and *Frappe*, and to a large extent the extensions mentioned earlier. We observed similar effects on the accuracy of ancestry recovery using *Structure* (Figure 5.7 in the Appendix). Likelihood-based methods are susceptible to the effects of sample selection bias since each individual is given equal weight in the sample a priori. This is a result of the fundamental assumption of many learning methods that the sample observed is representative of the underlying population distribution. Eigenanalysis, which also weighs each

individual equally a priori, also suffers from a similar problem as the likelihood-based method. The sensitivity of eigenanalysis to sample size variation and outliers is well known and has been reported by McVean [2009].

Our experiments used a simple two-population simulation scenario to examine the effects of sample selection bias on the accuracy of stratification. We used two populations that were easily separable and examined the data resulting after only a single generation of admixture. In reality, the demographic processes underlying the evolution of populations are much more complicated. In such scenarios, it is reasonable to expect that the stratification problem may be harder to resolve and would suffer from the effects of sample selection bias more severely.

Our experiments on data from the HGDP suggests that existing genetic datasets may also contain sample selection bias. Depending on the degree of such bias, the results obtained on these datasets may not be representative of the true evolutionary history and relationships of the underlying populations.

We proposed a resampling correction for sample selection bias using a mathematical framework we developed. The proposed correction requires knowledge of some auxiliary information that is correlated with the genotypes. For genetic datasets, geography provides one such criteria that is easy to acquire during data collection. Using this information, we proposed a correction that is easy to implement. We showed using simulation experiments and the HGDP data that such a correction is effective in practice and leads to more accurate results.

In our experiments, we either knew the nature of the bias (by design in the simulation experiments) or assumed that it was known. In general, correction of sample selection bias requires some domain-specific knowledge of the underlying bias. The accuracy of the correction method proposed will strongly depend on the relationship between the correction criteria and the genotypes. Corrections factors therefore may be dataset-specific. An alternative future direction for correcting sample selection bias would be to develop models of population structure that can also model the auxiliary factors, such as geography or language, that may determine the selec-

tion process responsible for the presence of the bias.

Appendix

An alternative evaluation metric

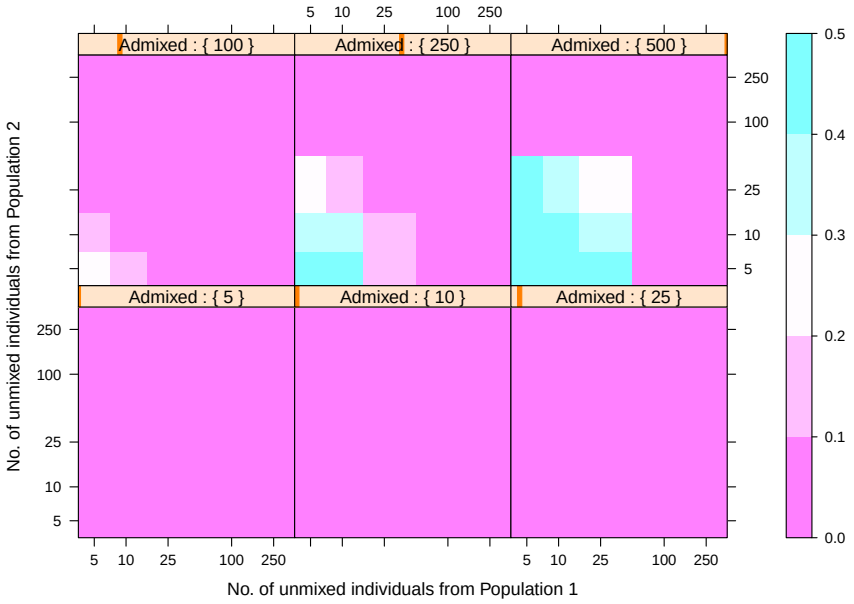


Figure 5.9: L1-norm error $|\theta^S_{true} - \theta^S_{infer}|$ between the true individual ancestry and the individual ancestry inferred using *Admixture* with K=2. Levelplots are drawn for different number of admixed individuals in the dataset.

An alternative metric for evaluating the recovery of individual ancestry inference is the mean L1-norm distance between the true and inferred ancestry vectors, i.e, $|\theta^S_{true} - \theta^S_{infer}|$, which measures the error in ancestry recovery. We find that this metric also shows similar behavior to the correlation measure of accuracy, with the error being low when the ratio of admixed individuals to unmixed individuals is less than 10 and high otherwise. Figure 5.9 shows the results using this error measure.

Chapter 6

Artificial selection for association

Disease association is the task of inferring genetic variants that contribute to disease risk or explain phenotypic diversity observed in inheritable traits. Traditional methods for genetic analysis of diseases used techniques such as linkage analysis of candidate markers or genes and quantitative trait locus (QTL) mapping using one marker and one phenotype at a time [Easton et al., 1993], followed by a correction for multiple hypothesis testing [Benjamini and Hochberg, 1995, Storey and Tibshirani, 2003]. Recently, methods have been developed that enhance power by allowing simultaneous analysis of multiple markers [Balding, 2006]. Methods such as eigenanalysis [Price et al., 2006] and regression [Cordell and Clayton, 2002] can perform simultaneous analysis of multiple markers for associations. Mixed models such as EMMA [Kang et al., 2008] extend the regression framework to model the association problem (with confounding variables) as a linear mixed model.

In this chapter, we propose an artificial selection setup for finding genetic associations. Artificial selection experiments belong to a class of experiments know as laboratory selection [Hill and Caballero, 1992], which can be used to answer questions about adaptations, trait associations, etc [Garland Jr and Garland, 2003]. In artificial selection experiments, individuals are chosen to propagate the next generation if they express particular values of a desired phenotypic trait. These experiments allow the experimenter more control over the selection experiment. We

show using simulated and semi-simulated data from artificial selection experiments that such methods enable better recovery of causal variants than conventional association techniques.

6.1 Proposed approach

The artificial selection experiment setup involves two sets of individuals, a selected group and a control group. The control group is a set of individuals on which no selection is performed. The selected group undergoes selection according to a regime of selection strength and consistency as chosen by the experimenter. As we described earlier, individuals from the selected group are chosen to reproduce to form the next generation if they express particular values of a phenotypic trait. For most traits, selection can be performed to obtain either high values of the trait or low values of the traits. Artificial selection experiments therefore often have two selected sets of individuals, one group selected for high values of the traits and the other selected for low values of the trait. To ensure that the experiment results are due to selection and not due to genetic drift, the experiment is often performed with more than one replicate.

The steps in an artificial selection experiment are:

1. Begin with an initial population of individuals as the current generation.
2. Measure the value of the phenotypic trait chosen for selection in all individuals in the current generation.
3. Individuals whose phenotype value matches a prespecified criterion for the phenotype (for example, trait value larger than an absolute or relative threshold) are chosen to be the parents for the individuals in the next generation.
4. The chosen parents are allowed to mate to produce a new generation of individuals. The number of individuals created is the same as that in the original population.
5. Repeat steps 2-4 with the new population.

Steps 2-5 are performed for the number of generations chosen by the experimenter.

We propose to set up an artificial selection experiment by breeding *Drosophila Melanogaster* for a trait of interest. We can then genotype (some of) the generations of individuals created during the artificial selection experiment. The sequenced genotypes and measured phenotypes can then be used for performing association between genotype and phenotype.

This approach has the advantage that it allows the experimenter control over the strength of selection. Low-frequency causual variants can therefore be enriched so that they can be detected statistically. In *Drosophila Melanogaster*, linkage disequilibrium decays to $r^2 = 0.2$ on average within 10 base pairs on autosomes [Mackay et al., 2012]. Therefore, causal variants can be found directly and not through association with a linked marker.

6.2 Large scale simulations

In our preliminary experiments [Shringarpure and Xing, 2012], we demonstrated how sparse regression methods can be used to perform association on data from an artificial selection experiment using genotypes from the initial (1^{st}) and final (20^{th}) generations. We simulated artificial selection experiments with 1200 individuals per generation and $100,000$ SNPs per individual. The recombination probability (between the ends of the chromosome) is set to 0.25 across all experiments. Individuals from a generation are chosen to be parents only if their phenotype is larger than the mean for the generation. For the experimental parameters, we varied the number of QTLs over $\{10, 20, 50\}$. The total heredity of the phenotype was varied over $\{0.01, 0.1, 0.3, 0.5\}$. The initial frequency of the QTLs was varied over $\{0.05, 0.1, 0.2\}$. Each experiment was repeated 20 times to obtain mean values of the F1 score of the methods at recovering the QTLs. To compute p-values for the sparse regression, we used the "screen and clean" (SC) method by Wasserman and Roeder [2007] and the adpative "multi-sample-split" (AMS) method proposed by Meinshausen et al. [2008]. We use the Cochran-Armitrage-Trend test (CATT) [Cochran, 1954] as a single-SNP test for association. For all methods, we used false discovery rate control (at the 5% level) using the method proposed by Benjamini and Yekutieli

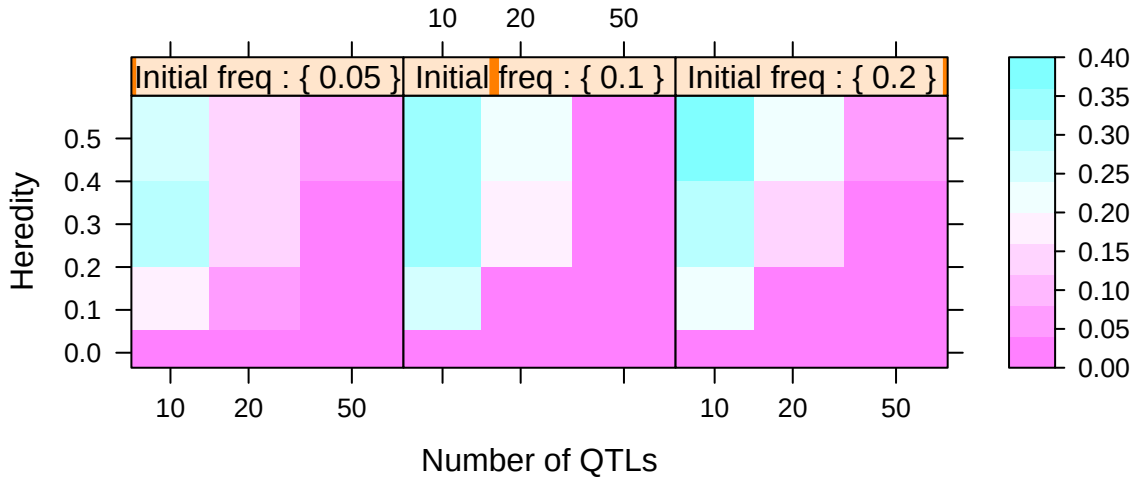


Figure 6.1: F1 performance of the adaptive multi-split method. The different panels are for the different initial QTL frequencies.

[2001].

Figures 6.1 and 6.2 shows the F1 results for the multi-split (AMS) and screen-and-clean (SC) regression methods respectively. The CATT produces F1 scores < 0.01 in almost all cases and therefore we do not plot its performance. Figure 6.3 shows the best performing method (multi-split) of the three when there is no artificial selection.

We find that conventional association (with no selection) performs well when the total heredity is high (0.5) and the number of loci is low (10), which is in line with our expectations of the scenarios in which genome-wide association studies are useful. In general though, performance at association is better under selection than without selection. AMS and SC perform well at recovering the QTLs with 20 QTLs even when total heredity is only 0.2. However, we find that when there are 50 QTLs, no method can recover QTLs reliably at the chosen sample size. We

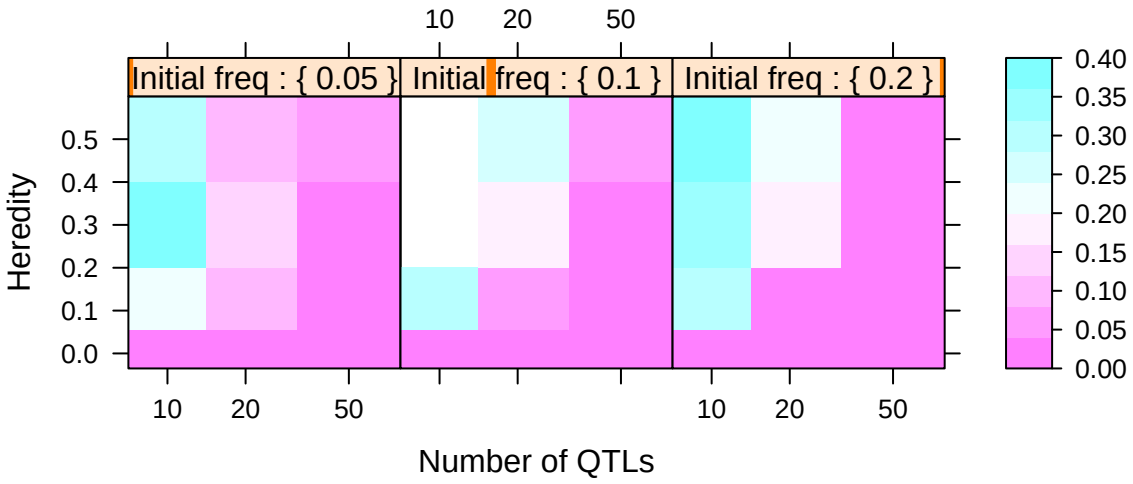


Figure 6.2: F1 performance of the screen-and-clean method. The different panels are for the different initial QTL frequencies.

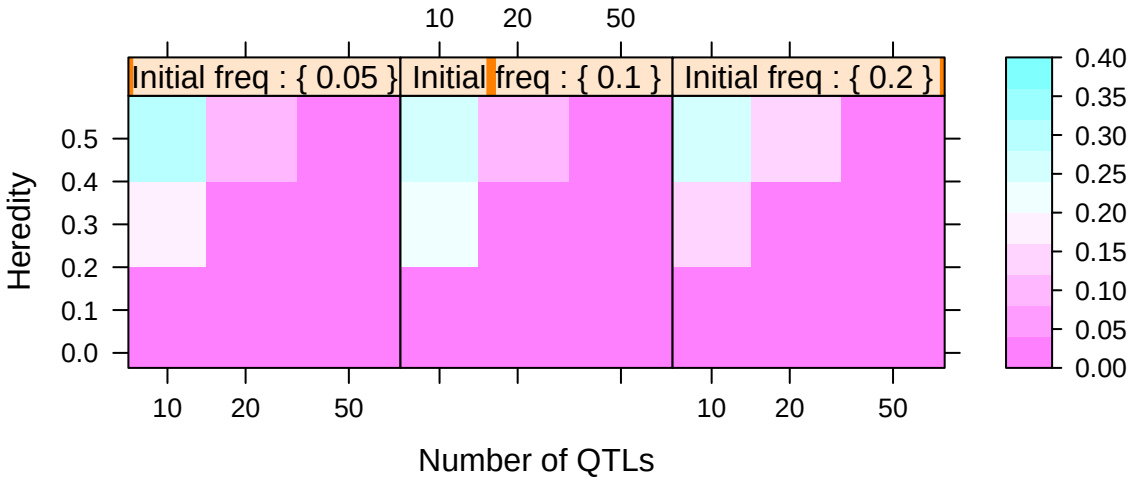


Figure 6.3: F1 performance under no selection. The different panels are for the different initial QTL frequencies.

also find that screen-and-clean performs slightly better than the adaptive multi-split method, but has higher variance.

6.2.1 Detecting epistasis

We simulated a phenotype with additive epistasis between two loci (with no individual effects). The heredity contribution of the epistasis term was varied over $\{0.001, 0.01, 0.1\}$ and was added to the parameters from the previous simulation. Epistasis terms can be added as a new covariate to the regression as the product $x_i * x_j$ (where x_i and x_j are the individual SNP genotypes at the i^{th} and j^{th} SNPs respectively). However, adding all possible pairs of SNPs is computational infeasible for the large number of SNPs we wish to include in the simulation or an actual study (indeed, we find that the fast exhaustive epsistasis testing method, BOOST [Wan et al., 2010] has running times larger than a day per run for the size of data we wish to simulate). Methods have therefore been proposed that suggest prioritizing pairs for inclusion based on their occurrence in SNP-SNP or gene-gene interaction networks [Lee and Xing, 2012]. We examine the power of sparse regression methods at recovering the true epistatic interaction term. For this experiment, we included 10,000 interaction terms in the sparse regression along with the individual SNPs. Figures 6.4 and 6.5 show the power of the AMS and SC methods at recovering the epistatic interaction. Figure 6.6 shows the power at recovering the epistatic interaction when there is no selection.

As before, we find that power at recovering epistasis is much higher when there is selection. We also find the screen-and-clean performs better than the adaptive multi-split method. We can observe the effect of using a sparse regression which considers all the association covariates simultaneously in that the power varies not only as a function of the heredity contribution due to the epistatic term, but also as a function of the total heredity of the trait.

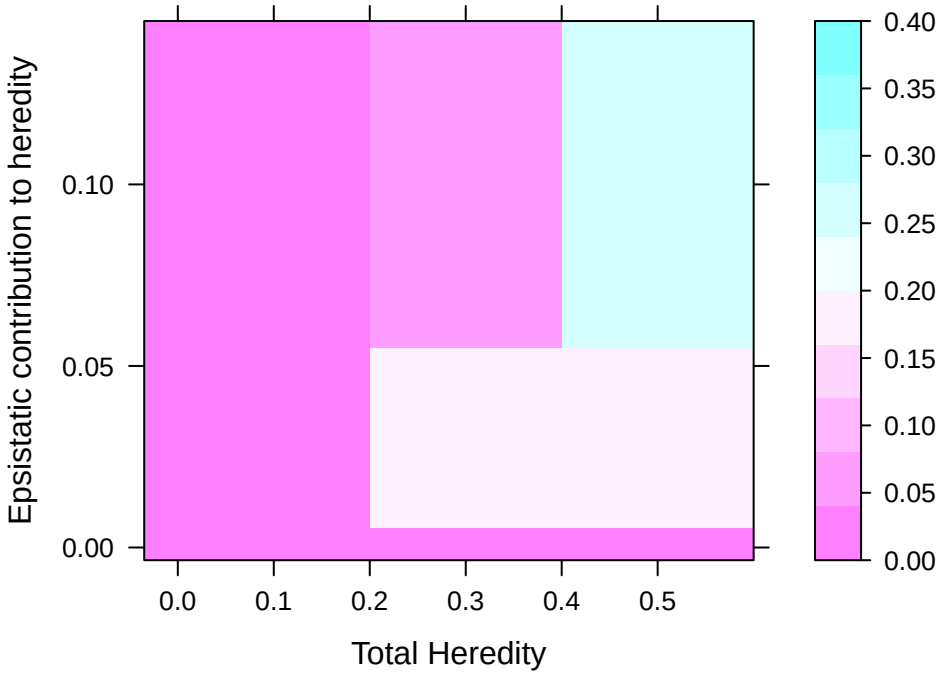


Figure 6.4: Power of the multi-split method at recovering the epistatic interaction. The power is plotted as function of the total heredity of the trait and the heredity contributed by the epistatic interaction term.

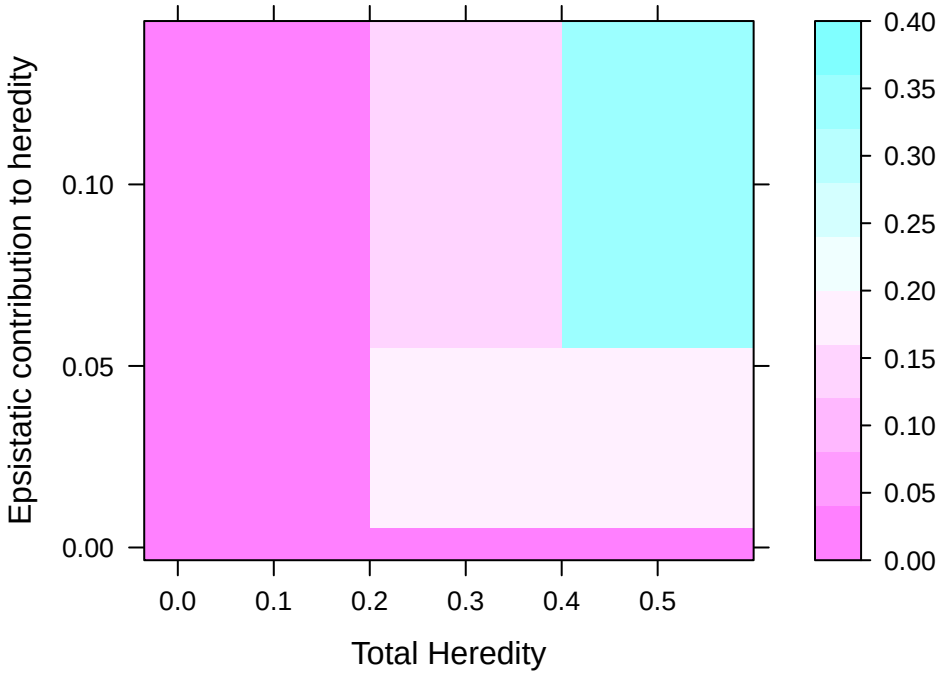


Figure 6.5: Power of the screen-and-clean method at recovering the epistatic interaction. The power is plotted as function of the total heredity of the trait and the heredity contributed by the epistatic interaction term.

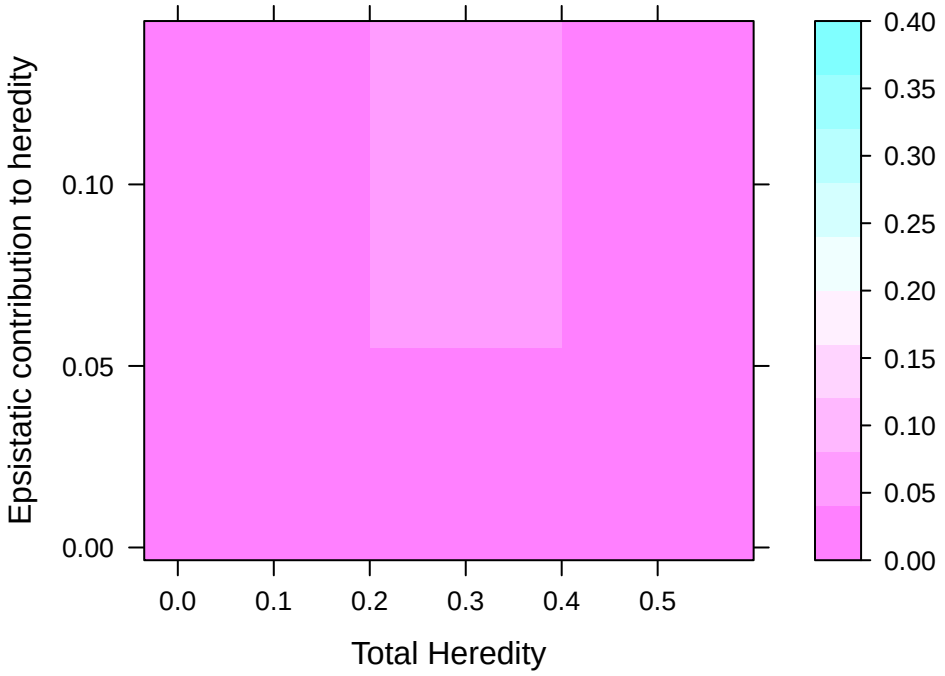


Figure 6.6: Power at recovering the epistatic interaction when there is no selection. The power is plotted as function of the total heredity of the trait and the heredity contributed by the epistatic interaction term.

6.3 Analysis of data from an artificial selection experiment on *Drosophila Melanogaster*

Burke et al. [2010] performed a selection experiment on *Drosophila Melanogaster* for accelerated development. Flies in the selected populations develop from egg to adult 20% faster than flies of ancestral control populations. The resequencing data from these populations contains allele frequencies for 688,520 SNPs in the control and selected populations. Using this data, they identified 30 SNPs (out of 662 potential candidate SNPs) that show significant allele frequency differences between the selected populations and the control populations. These 30 SNPs were genotyped in 35 females in each of the populations (a total of 175 selected individuals and 175 control individuals) using cleaveage amplified polymorphic sequence (CAPS) techniques. The data we obtained therefore consists of 350 individuals genotyped at 30 SNP loci. 175 of the 350 individuals have undergone selection for accelarated development and 175 are control individuals.

6.3.1 Experiment setup

We use the genotype data from Burke et al. [2010] to verify that the sparse regression method we proposed works on data from an existing artificial selection experiment. Since we do not have genotype data for the other SNPs in the dataset, we cannot perform sparse regression to find how well the predictions from regression agree with the 30 known SNPs. We therefore use an alternate way of verifying the accuracy of the regression by simulating neutral SNPs.

We generated N (for different values of N) neutral SNPs of varying minor allele frequency (MAF) and constructed genotypes in the form of minor allele counts at each SNP for the 350 individuals according to the MAF value at the SNP. We then combined the genotypes at the 30 candidate SNPs and N artificial SNPs and ran the sparse regression, using the individuals' status as control (0) or selected (1) as the response variable. The sparse regression was set to choose

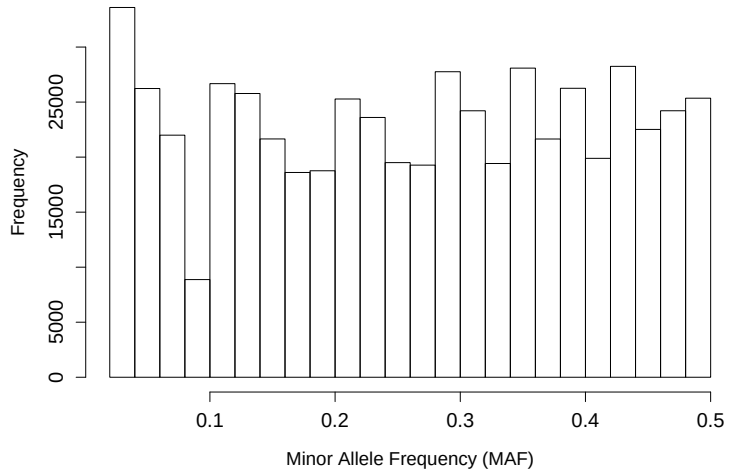


Figure 6.7: Histogram of MAF values for the 688,520 SNPs in the dataset

the top 30 most relevant SNPs out of the N+30 total SNPs. The number of the original 30 SNPs chosen by the regression then gives us an estimate of the accuracy of the regression. We can also examine which of the original 30 SNPs are chosen by the regression. Repeating the experiment multiple times for a particular value of N allows us to get a statistically meaningful estimate of the result.

To see the effect of changing the number of non-causal SNPs, we varied N from 1 to 1 million in powers of ten. From the data, we can also observe that the distribution of minor allele frequency (MAF) is almost uniform over (0,0.5], as seen in Figure 6.7. We generated artificial SNPs so that the distribution of MAFs matched this observed distribution. For statistically meaningful estimates, 50 runs of simulation and regression were set up for each value of N.

6.3.2 Results

For each setting of N, we observed how many times the 30 candidate SNPs were chosen by the regression in the 50 runs. Figure 6.8 shows the results.

From the figure, we can see that as the number of simulated SNPs increases, fewer of the 30

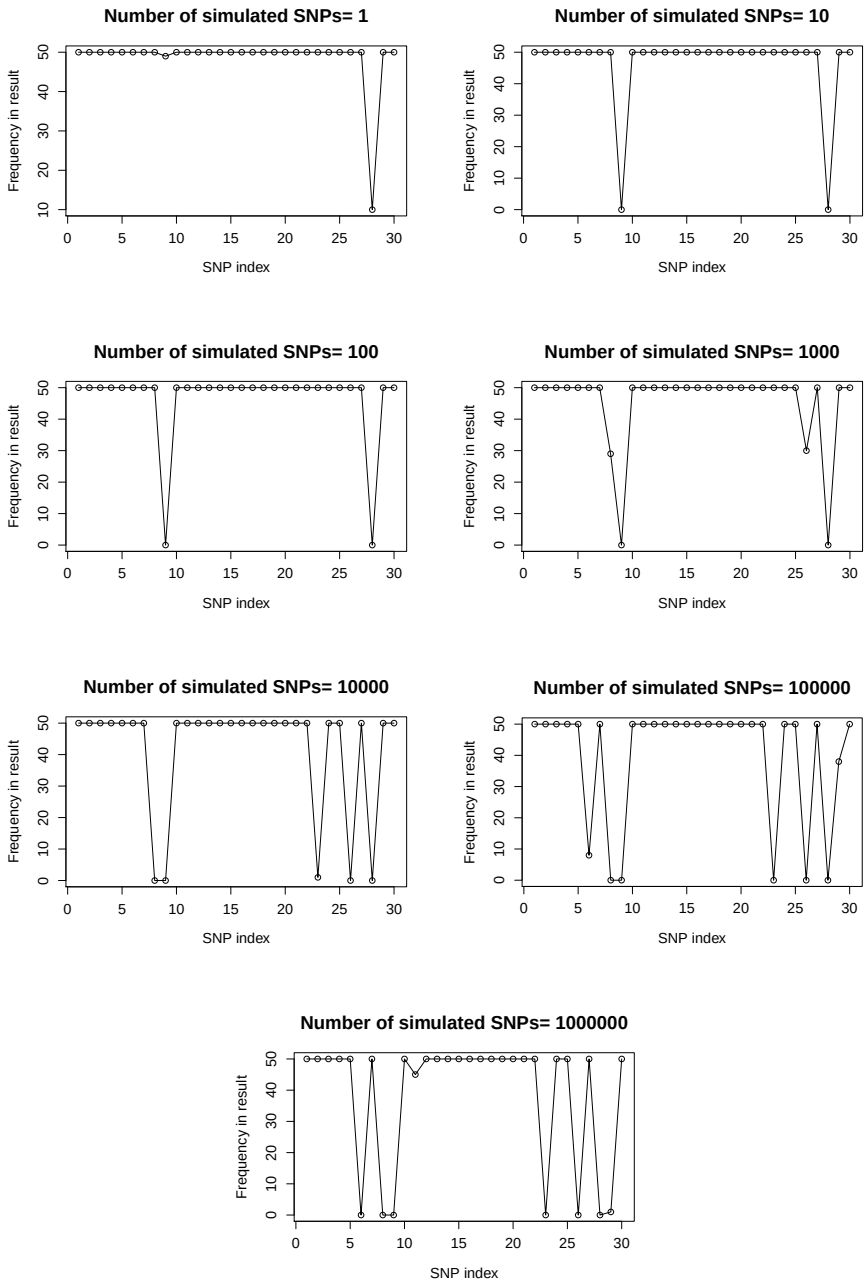


Figure 6.8: Simulation results. On the X-axis are the indices of the 30 candidate SNPs and the Y-axis shows how many times they are chosen by the regression in 50 runs. As the number of artificial SNPs increases, some SNPs out of the 30 candidates stop being chosen by the regression. However, even when 1 million artificial SNPs are added to the data, 23 of the original 30 SNPs show a strong signal and are chosen.

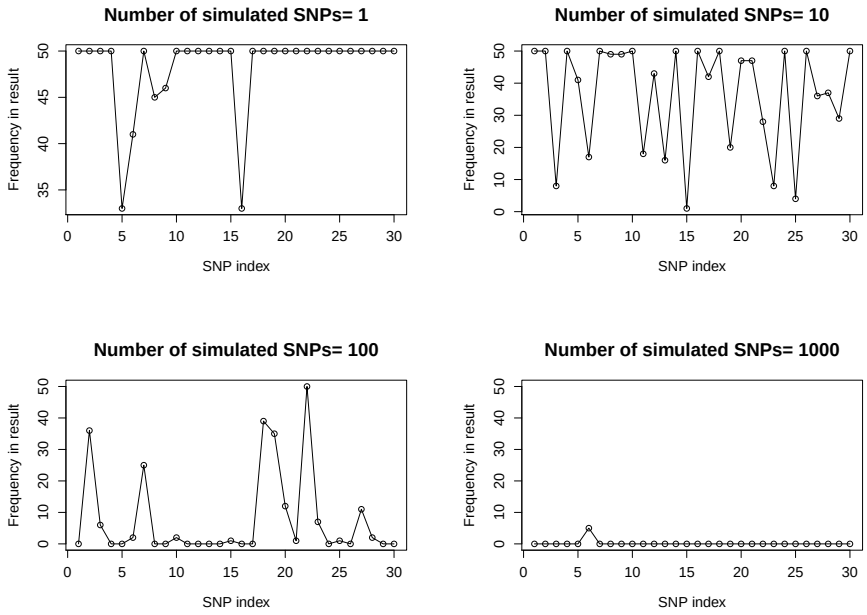


Figure 6.9: Simulation results using only control populations. On the X-axis are the indices of the 30 candidate SNPs and the Y-axis shows how many times they are chosen by the regression in 50 runs. As the number of artificial SNPs increases, the regression method is unable to pick any of the original 30 candidate SNPs.

candidate SNPs are chosen by the regression. The set of SNPs chosen by the regression decreases almost monotonically, i.e, once a SNP is not chosen by regression, it is never chosen again as the number of simulated SNPs increases. However, even when 1 million artificial SNPs are added to the data, 23 of the original 30 SNPs show a strong signal and are chosen by the regression. This suggests that the regression method does well at picking loci that have undergone selection.

To ensure that this is not an artifact, we performed the simulation using only the individuals from the control populations. Figure 6.9 shows the results for N upto 1000. We can see that even at N=1000, almost none of the original 30 candidate SNPs can be picked up by the regression.

Since the simulation using only control populations has only half the number of individuals as the original populations, it is helpful to see what effect sample size has when analyzing the combined populations. To test this, we used only half the individuals from the control and selected populaitons and repeated the experiments. The results are shown in Figure 6.10. We

can see that there is a significant impact on performance at recovering the original 30 SNPs. With 1 million simulated SNPs, only 15 of the original 30 SNPs are chosen by the regression.

6.4 Discussion

We proposed that artificial selection can be used in conjunction with sparse regression techniques to perform association better. Our simulation experiments show that data from artificial selection experiments enables better recovery of the true causual variants than a conventional association setup with no selection. For traits with a small number of QTLS and high heredity, association without selection performs better than our proposed approach but its performance degrades quickly as the number of QTLs increases and the heredity of the trait reduces to moderate or low amounts. We also find that the proposed method performs better than conventional association when trying to recover additive epistatic effects.

The analysis of genotype data from an artificial selection on *Drosophila Melanogaster* by Burke et al. [2010] with the simulated noise SNPs show that the regression method does well at identifying the original 30 candidate SNPs (23 of the 30 are picked up even when there are 1 million non-causal SNPs). On the other hand, using only the control populations along with simulated genotypes does not enable us to identify any of the candidate SNPs even when there are only 1000 artificial SNPs. This shows that our method works because it can identify the difference between control individuals and selected individuals. It is important to note that the data from Burke et al. [2010] is from an artificial selection experiment that lasted 600 generations. The differentiation between the control and selected populations is therefore quite large, which makes the problem easier.

Another phenomenon that we can observe in these experiments is the effect of sample size. Using only control populations reduces the number of available samples by half, which affects the recovery of the 30 candidate SNPs. If we halve the total number of individuals from the complete dataset and perform regression on the resulting dataset, we find that the accuracy of

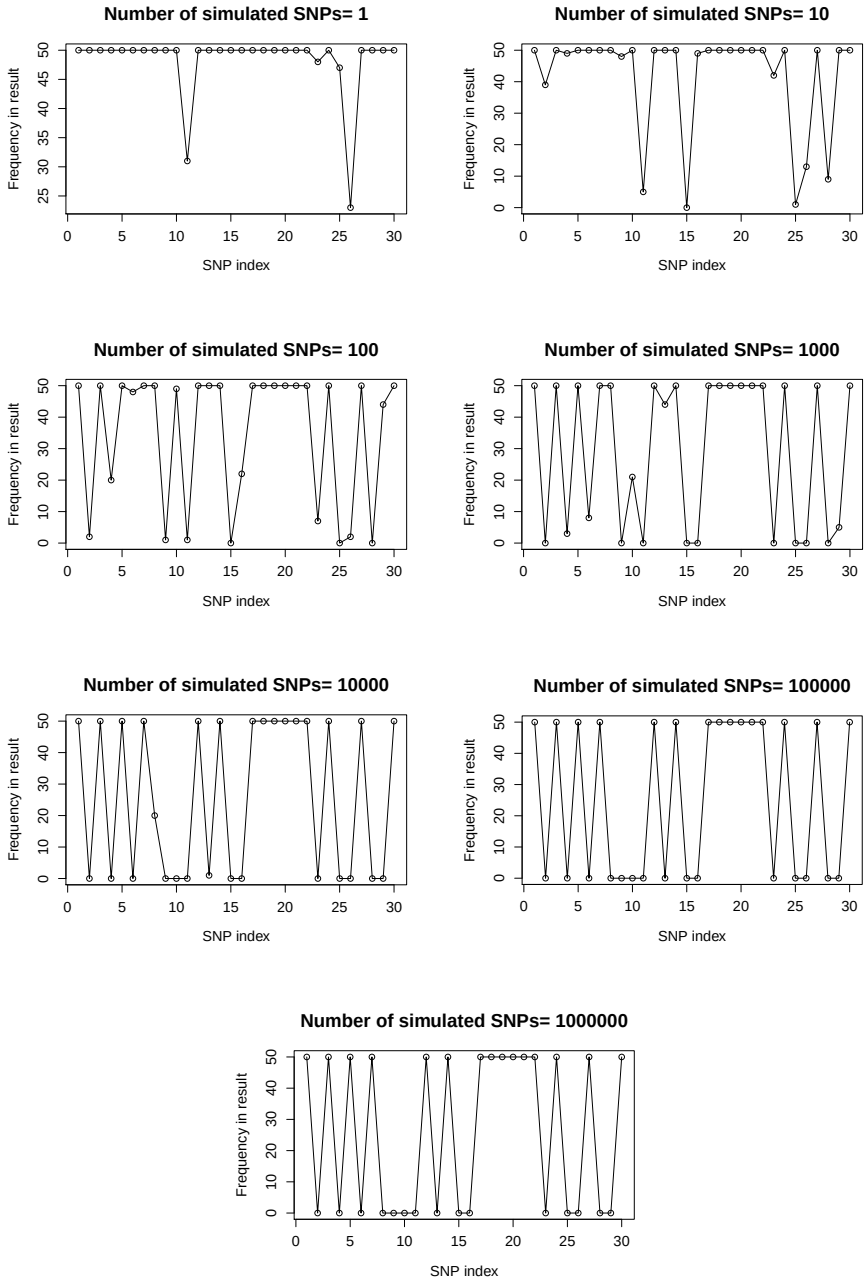


Figure 6.10: Simulation results using half the number of individuals. On the X-axis are the indices of the 30 candidate SNPs and the Y-axis shows how many times they are chosen by the regression in 50 runs. As the number of artificial SNPs increases, some SNPs out of the 30 candidates stop being chosen by the regression. When 1 million artificial SNPs are added to the data, 15 of the original 30 SNPs show a strong signal and are chosen.

the regression is affected noticeably and the number of candidates recovered by the regression drops to about 15. A sample size of 300-400 therefore seems to be necessary for recovering associations. However, all the p-value computation methods we examine in our analysis require splitting of data into multiple parts. Therefore, to obtain reliable p-values from such analyses, sample sizes close to 1000 may be more desirable. At these sample sizes, our proposed approach can recover QTLs with some success for a moderate number of QTLs (20) and moderate values of heredity (0.1). For the method to work reliably for a larger number of causal variants, even larger sample sizes would be necessary.

Artificial selection approaches have advantages over conventional association that are important for performing association. They provide the experimenter with greater control on the expression of the trait of interest in the selected population. Using model organisms such as *Drosophila Melanogaster* which show weak linkage disequilibrium allows more accurate localization of the causal variants. Burke et al. [2010] demonstrate that artificial selection experiments can be used to determine causal variants for accelarated development in *Drosophila Melanogaster*. With larger sample sizes, artificial selection experiments are likely to provide an effective way of performing association on complex traits.

Chapter 7

Conclusions and Future work

In this thesis, we have developed methods to improve our understanding of genetic variation in populations - population structure detection and disease association. Both these problems have been extensively studied in the literature and our attempts aim to extend our understanding using efficient statistical methods that can model the evolutionary processes that shape genetic variation.

In Chapter 3, we presented a hierarchical Bayesian model, *mStruct* [Shringarpure and Xing, 2009], that can model admixing of populations along with allele mutations. We developed an efficient inference algorithm for the model using variational inference. Our simulations showed that modeling mutations allows us to model ancestries more accurately than a model which only takes admixture into account. Analysis of data from the Human Genome Diversity project showed that *mStruct* allows us to model similarities and differences between populations in a meaningful way. It also enables the study of the accumulated mutation in populations, which can be used to qualitatively estimate the age of populations.

The *mStruct* method developed in Chapter 3, like most other methods used to analyze population structure, requires the user to specify the number of ancestral populations that contribute to the given sample of individuals. This requires prior knowledge of the evolutionary history of the sample and may not always be possible or desirable. Chapter 4 addressed the problem of choos-

ing the number of ancestral populations using a non-paramatric Bayesian model, *StructHDP*, in a data-dependent manner [Shringarpure et al., 2011].

Through experiments on real and simulated data, Chapter 5 showed that biased sampling can affect the results of ancestry inference using likelihood-based methods. We developed a mathematical framework for modeling sample selection bias and also proposed a correction that is easy to implement and effective in practice. This is the first attempt to address sample selection bias in an unsupervised clustering setting. As more datasets of genetic sequences become available for analyses and are combined in an attempt to improve the accuracy and resolution of ancestry inference, we believe that sample selection bias will become an important concern that will need to be addressed. A useful extension of existing methods that could account for sample selection bias would be to develop methods that can incorporate weights for samples directly into the models.

Another problem that must be addressed with larger datasets is that of scalability. The methods we have proposed for population structure detection, while efficient for smaller datasets, do not scale well to genotypes consisting of hundreds of thousands or millions of SNPs. In such cases, development of efficient inference algorithms requires better optimization techniques, such as the quasi-Newton techniques used in *Admixture* [Alexander and Lange, 2011, Alexander et al., 2009]. We also note that using genotype data with many SNPs is likely to violate the assumptions of unlinked loci made in *mStruct* and *StructHDP*.

We have made an attempt to incorporate evolutionary processes affecting genetic variation into our models. Our experiments with modeling mutation in *mStruct* show that this can improve the accuracy of ancestry inference. It is therefore natural to expect that extensions incorporating other evolutionary process such as recombination and selection into the modeling will improve ancestry inference. An important caveat in making more expressive models is that such models have high complexity and may not produce biologically meaningful results in the absence of enough data and constraints on the model. For instance, the ancestral populations inferred using

these models are mathematical entities and may have no true historical counterpart. Recently, however, many new sequencing projects have genotyped ancient individuals from various geographical regions, such as the Neandarthals [Green et al., 2010], the Tyrolean Iceman [Ermini et al., 2008], the Denisovans [Reich et al., 2010], Australian aborigines [Rasmussen et al., 2011]. Data from these genome sequences may allow the development of models where the ancestral populations can be constrained to be similar to these putative ancestral populations.

Chapter 6 examined the problem of disease association through a novel approach. We proposed that genotype data from artificial selection experiments enables us to perform association with better power than conventional approaches. We proposed the use of sparse regression methods to perform the association efficiently, and demonstrated the validity of the claims using simulated data and semi-simulated data from an artificial selection experiment on *Drosophila Melanogaster*. Our experiments also suggest that sample size is an important factor in determining the accurate recovery QTLs when a trait is affected by a large number of loci with small individual effects. Similiar observations have been made in a previous study by Yang et al. [2010] in the context of human height. Another important problem of future interest would be to develop ways of determing the statistical significance of the predictions (made by the sparse regression methods) that can make effective use of available data. Current methods [Meinshausen et al., 2008, Wasserman and Roeder, 2007] require the data to be split into two or more parts, which greatly reduces the power of association methods when dealing with a very large number of loci.

Bibliography

H Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. ISSN 00189286. doi: 10.1109/TAC.1974.1100705. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1100705. 4.1

David H Alexander and Kenneth Lange. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC bioinformatics*, 12(1):246, January 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-246. URL http://www.biomedcentral.com/1471-2105/12/246. 7

David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.094052.109. 7

D.M. Altshuler, E.S. Lander, L. Ambrogio, T. Bloom, K. Cibulskis, T.J. Fennell, S.B. Gabriel, D.B. Jaffe, E. Shefler, , and C.L. others Sougnez. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, October 2010. ISSN 1476-4687. doi: 10.1038/nature09534. URL http://dx.doi.org/10.1038/nature09534http://www.pubmedcentral.nih.gov/articlerender.

fcgi?artid=3042601&tool=pmcentrez&rendertype=abstract. 1

E C Anderson and E A Thompson. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, 160(3):1217–1229, 2002. 2.2, 3.1

Antonis C Antoniou, Amanda B Spurdle, Olga M Sinilnikova, Sue Healey, Karen A Pooley, Rita K Schmutzler, Beatrix Versmold, Christoph Engel, Alfons Meindl, Norbert Arnold, Wera Hofmann, Christian Sutter, Dieter Niederacher, Helmut Deissler, Trinidad Caldes, Kati Kämpjärvi, Heli Nevanlinna, Jacques Simard, Jonathan Beesley, Xiaoqing Chen, Susan L Neuhausen, Timothy R Rebbeck, Theresa Wagner, Henry T Lynch, Claudine Isaacs, Jeffrey Weitzel, Patricia A Ganz, Mary B Daly, Gail Tomlinson, Olufunmilayo I Olopade, Joanne L Blum, Fergus J Couch, Paolo Peterlongo, Siranoush Manoukian, Monica Barile, Paolo Radice, Csilla I Szabo, Lutecia H Mateus Pereira, Mark H Greene, Gad Rennert, Flavio Lejbkowicz, Ofra Barnett-Griness, Irene L Andrulis, Hilmi Ozcelik, Anne-Marie Gerdes, Maria A Caligo, Yael Laitman, Bella Kaufman, Roni Milgrom, Eitan Friedman, Susan M Domchek, Katherine L Nathanson, Ana Osorio, Gemma Llort, Roger L Milne, Javier Benítez, Ute Hamann, Frans B L Hogervorst, Peggy Manders, Marjolijn J L Ligtenberg, Ans M W van den Ouweland, Susan Peock, Margaret Cook, Radka Platte, D Gareth Evans, Rosalind Eeles, Gabriella Pichert, Carol Chu, Diana Eccles, Rosemarie Davidson, Fiona Douglas, Andrew K Godwin, Laure Barjhoux, Sylvie Mazoyer, Hagay Sobol, Violaine Bourdon, François Eisinger, Agnès Chompret, Corinne Capoulade, Brigitte Bressac-de Paillerets, Gilbert M Lenoir, Marion Gauthier-Villars, Claude Houdayer, Dominique Stoppa-Lyonnet, Georgia Chenevix-Trench, and Douglas F Easton. Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. *American Journal of Human Genetics*, 82(4):937–948, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18355772. 1

David J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, October 2006. ISSN 1471-0056. doi: 10.1038/nrg1916.

URL http://www.ncbi.nlm.nih.gov/pubmed/16983374. 2.4, 6

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, 57(1):289–300, 1995. ISSN 00359246. doi: 10.2307/2346101. URL http://www.jstor.org/stable/2346101. 6

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. ISSN 00905364. doi: 10.2307/2674075. URL http://www.jstor.org/stable/2674075. 6.2

Marnie Bertolet. *To weight or not to weight? Incorporating sampling designs into model-based analyses*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA - 15213, 2008. 5.2.2

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, May 2003. ISSN 1532-4435. doi: 10.1162/jmlr.2003.3.4-5.993. 2.3, 3.1, 3.4

A M Bowcock, A Ruiz-Linares, J Tomfohrde, E Minch, J R Kidd, and L L Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455–457, 1994. ISSN 00280836. URL http://www.ncbi.nlm.nih.gov/pubmed/7510853. 1, 2.2, 3.2.3, 3.5.2

Molly K. Burke, Joseph P. Dunham, Parvin Shahrestani, Kevin R. Thornton, Michael R. Rose, and Anthony D. Long. Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature*, 467(7315):587–590, September 2010. ISSN 0028-0836. doi: 10.1038/nature09352. URL http://dx.doi.org/10.1038/nature09352. 6.3, 6.3.1, 6.4

H M Cann, C de Toma, L Cazes, M F Legrand, V Morel, L Piouffre, J Bodmer, W F Bodmer, B Bonne-Tamir, A Cambon-Thomsen, and Others. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002. 1, 3.5.2

L L Cavalli-Sforza, P Menozzi, and A Piazza. *The history and geography of human genes*. Princeton Univ Pr, 1994. 1, 2.2, 4.6

L Luca Cavalli-Sforza. The Human Genome Diversity Project: past, present and future. *Nature Reviews Genetics*, 6(4):333–340, 2005. URL http://www.nature.com/doifinder/10.1038/nrg1579. 1, 3.5.2, 5.1

William G Cochran. Some methods for strengthening the common chi-square tests. *Biometrics*, 10(4):417–451, 1954. 6.2

F S Collins, M S Guyer, and A Charkravarti. Variations on a theme: cataloging human DNA sequence variation. *Science New York NY*, 278(5343):1580–1581, 1997. URL http://www.ncbi.nlm.nih.gov/pubmed/9411782.

D F Conrad, M Jakobsson, G Coop, X Wen, J D Wall, N A Rosenberg, and J K Pritchard. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *NatureGenetics*, 38:1251–1260, 2006. 1, 2.2, 3, 3.1, 3.5.2

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14 , 000 cases of seven common diseases and. *Nature*, 447(June), 2007. doi: 10.1038/nature05911. 1

J Corander, P Waldmann, and M J Sillanpaa. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1):367–374, 2003. 2.2, 3.1

Jukka Corander and Pekka Marttinen. Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, 15(10):2833–2843, 2006. URL http://www.ncbi.nlm.nih.gov/pubmed/16911204.

Heather J Cordell and David G Clayton. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *The American Journal of Human Genetics*, 70(1):124–141, 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/11719900. 2.4, 6

Heather J Cordell and David G Clayton. Genetic association studies. *Lance*, 366(9491):1121–

1131, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/16182901. 1, 2.4

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample Selection Bias Correction Theory. *Algorithmic Learning Theory*, 5254:16, 2008. URL http://arxiv.org/abs/0805.2775. 5.2.2

I Davidson and B Zadrozny. An improved categorization of classifiers sensitivity on sample selection bias. *Fifth IEEE International Conference on Data Mining ICDM05*, pages 605–608, 2005. doi: 10.1109/ICDM.2005.24. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1565737. 5.2.2

W Dietrich, H Katz, S E Lincoln, H S Shin, J Friedman, N L Dracopoli, and E S Lander. A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics*, 131(2):423–447, 1992. 3.2.3

Richard M Durbin, David L Altshuler, Gonçalo R Abecasis, David R Bentley, Aravinda Chakravarti, Andrew G Clark, Francis S Collins, Francisco M De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B Gabriel, Richard A Gibbs, Bartha M Knoppers, Eric S Lander, Hans Lehrach, Elaine R Mardis, Gil A McVean, Debbie A Nickerson, Leena Peltonen, Alan J Schafer, Stephen T Sherry, Jun Wang, Richard K Wilson, David Deiros, Mike Metzker, Donna Muzny, Jeff Reid, David Wheeler, Jingxiang Li, Min Jian, Guoqing Li, Ruiqiang Li, Huiqing Liang, Geng Tian, Bo Wang, Jian Wang, Wei Wang, Huanming Yang, Xiuqing Zhang, Huisong Zheng, Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J Fennell, David B Jaffe, Erica Shefler, Carrie L Sougnez, Niall Gormley, Sean Humphray, Zoya Kingsbury, Paula Koko-Gonzales, Jennifer Stone, Kevin J McKernan, Gina L Costa, Jeffry K Ichikawa, Clarence C Lee, Ralf Sudbrak, Tatiana A Borodina, Andreas Dahl, Alexey N Davydov, Peter Marquardt, Florian Mertes, Wilfiried Nietfeld, Philip Rosenstiel, Stefan Schreiber, Aleksey V Soldatov, Bernd Timmermann, Marius Tolzmann, Jason Affourtit, Dana Ashworth, Said Attiya, Melissa Bachorski, Eli Buglione, Adam Burke, Amanda Caprio, Christopher Celone, Shauna Clark, David Conners, Brian Desany, Lisa Gu, Lorri

Guccione, Kalvin Kao, Andrew Kebbel, Jennifer Knowlton, Matthew Labrecque, Louise McDade, Craig Mealmaker, Melissa Minderman, Anne Nawrocki, Faheem Niazi, Kristen Pareja, Ravi Ramenani, David Riches, Wanmin Song, Cynthia Turcotte, Shally Wang, David Dooling, Lucinda Fulton, Robert Fulton, George Weinstock, John Burton, David M Carter, Carol Churcher, Alison Coffey, Anthony Cox, Aarno Palotie, Michael Quail, Tom Skelly, James Stalker, Harold P Swerdlow, Daniel Turner, Anniek De Witte, Shane Giles, Matthew Bainbridge, Danny Challis, Aniko Sabo, Fuli Yu, Jin Yu, Xiaodong Fang, Xiaosen Guo, Yingrui Li, Ruibang Luo, Shuaishuai Tai, Honglong Wu, Hancheng Zheng, Xiaole Zheng, Yan Zhou, Gabor T Marth, Erik P Garrison, Weichun Huang, Amit Indap, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Aaron R Quinlan, Chip Stewart, Michael P Stromberg, Alistair N Ward, Jiantao Wu, Charles Lee, Ryan E Mills, Xinghua Shi, Mark J Daly, Mark A DePristo, Aaron D Ball, Eric Banks, Brian L Browning, Kiran V Garimella, Sharon R Grossman, Robert E Handsaker, Matt Hanna, Chris Hartl, Andrew M Kernytsky, Joshua M Korn, Heng Li, Jared R Maguire, Steven A McCarroll, Aaron McKenna, James C Nemesh, Anthony A Philippakis, Ryan E Poplin, Alkes Price, Manuel A Rivas, Pardis C Sabeti, Stephen F Schaffner, Ilya A Shlyakhter, David N Cooper, Edward V Ball, Matthew Mort, Andrew D Phillips, Peter D Stenson, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C Yoon, Carlos D Bustamante, Adam Boyko, Jeremiah Degenhardt, Simon Gravel, Ryan N Gutenkunst, Mark Kaganovich, Alon Keinan, Phil Lacroute, Xin Ma, Andy Reynolds, Laura Clarke, Fiona Cunningham, Javier Herrero, Stephen Keenen, Eugene Kulesha, Rasko Leinonen, William M McLaren, Rajesh Radhakrishnan, Richard E Smith, Vadim Zalunin, Xiangqun Zheng-Bradley, Jan O Korbel, Adrian M Stütz, Markus Bauer, R Keira Cheetham, Tony Cox, Michael Eberle, Terena James, Scott Kahn, Lisa Murray, Kai Ye, Yutao Fu, Fiona C L Hyland, Jonathan M Manning, Stephen F McLaughlin, Heather E Peckham, Onur Sakarya, Yongming A Sun, Eric F Tsung, Mark A Batzer, Miriam K Konkel, Jerilyn A Walker, Marcus W Albrecht, Vyacheslav S Amstislavskiy, Ralf Herwig, Dimitri V Parkhomchuk, Richa Agarwala, Hoda M Khouri, Alek-

sandr O Morgulis, Justin E Paschall, Lon D Phan, Kirill E Rotmistrovsky, Robert D Sanders, Martin F Shumway, Chunlin Xiao, Adam Auton, Zamin Iqbal, Gerton Lunter, Jonathan L Marchini, Loukas Moutsianas, Simon Myers, Afidalina Tumian, James Knight, Roger Winer, David W Craig, Steve M Beckstrom-Sternberg, Alexis Christoforides, Ahmet A Kurdoglu, John V Pearson, Shripad A Sinari, Waibhav D Tembe, David Haussler, Angie S Hinrichs, Sol J Katzman, Andrew Kern, Robert M Kuhn, Molly Przeworski, Ryan D Hernandez, Bryan Howie, Joanna L Kelley, S Cord Melton, Yun Li, Paul Anderson, Tom Blackwell, Wei Chen, William O Cookson, Jun Ding, Hyun Min Kang, Mark Lathrop, Liming Liang, Miriam F Moffatt, Paul Scheet, Carlo Sidore, Matthew Snyder, Xiaowei Zhan, Sebastian Zöllner, Philip Awadalla, Ferran Casals, Youssef Idaghdour, John Keebler, Eric A Stone, Martine Zilversmit, Lynn Jorde, Jinchuan Xing, Evan E Eichler, Gozde Aksay, Can Alkan, Iman Hajirasouliha, Fereydoun Hormozdiari, Jeffrey M Kidd, S Cenk Sahinalp, Peter H Sudmant, Ken Chen, Asif Chinwalla, Li Ding, Daniel C Koboldt, Mike D McLellan, John W Wallis, Michael C Wendl, Qunyuan Zhang, Cornelis A Albers, Qasim Ayub, Senduran Balasubramaniam, Jeffrey C Barrett, Yuan Chen, Donald F Conrad, Petr Danecek, Emmanouil T Dermitzakis, Min Hu, Ni Huang, Matt E Hurles, Hanjun Jin, Luke Jostins, Thomas M Keane, Si Quang Le, Sarah Lindsay, Quan Long, Daniel G MacArthur, Stephen B Montgomery, Leopold Parts. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010. URL http://www.nature.com/doifinder/10.1038/nature09534. 2.1

D F Easton, D T Bishop, D Ford, and G P Crockford. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *The American Journal of Human Genetics*, 52(4):678–701, 1993. 1, 2.4, 6

Rosalind A Eeles, Zsofia Kote-Jarai, Graham G Giles, Ali Amin Al Olama, Michelle Guy, Sarah K Jugurnauth, Shani Mulholland, Daniel A Leongamornlert, Stephen M Edwards, Jonathan Morrison, Helen I Field, Melissa C Southey, Gianluca Severi, Jenny L Donovan, Freddie C Hamdy, David P Dearnaley, Kenneth R Muir, Charmaine Smith, Melisa Bagnato,

Audrey T Ardern-Jones, Amanda L Hall, Lynne T O'Brien, Beatrice N Gehr-Swain, Rosemary A Wilkinson, Angie Cox, Sarah Lewis, Paul M Brown, Sameer G Jhavar, Malgorzata Tymrakiewicz, Artitaya Lophatananon, Sarah L Bryant, Alan Horwich, Robert A Huddart, Vincent S Khoo, Christopher C Parker, Christopher J Woodhouse, Alan Thompson, Tim Christmas, Chris Ogden, Cyril Fisher, Charles Jamieson, Colin S Cooper, Dallas R English, John L Hopper, David E Neal, and Douglas F Easton. Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genetics*, 40(3):316–321, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18264097. 1

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. ISSN 00905364. doi: 10.1214/009053604000000067. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.aos/1083178935/.

Barbara E Engelhardt and Matthew Stephens. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS genetics*, 6(9): 12, September 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001117. URL http://dx.plos.org/10.1371/journal.pgen.1001117.

L. Ermini, C. Olivieri, E. Rizzi, G. Corti, R. Bonnal, P. Soares, S. Luciani, I. Marota, G. De Bellis, M.B. Richards, et al. Complete mitochondrial genome sequence of the tyrolean iceman. *Current Biology*, 18(21):1687–1693, 2008. 7

E Erosheva, SE Fienberg, and J Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(90001):5220–5227, 2004. 2.3

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577, June 1995. ISSN 01621459. doi: 10.2307/2291069. URL http://www.jstor.org/stable/2291069?origin=crossref. 4.4.1

G Evanno, S Regnaut, and J Goudet. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, 14(8):2611–2620, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15969739.

L Excoffier and G Hamilton. Comment on genetic structure of human populations. *Science*, 300 (5627):1877, 2003. 3.1

L Excoffier, J Novembre, and S Schneider. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered*, 91(6):506–509, 2000.

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164 (4):1567–1587, August 2003. ISSN 0016-6731. URL http://www.ncbi.nlm.nih.gov/pubmed/12930761. 2.3, 2.3.1, 2.3.2, 3.6, 4.6

J Felsenstein and Others. *Inferring phylogenies*. Sinauer Associates Sunderland, Mass., USA, 2004. 3.6

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973. ISSN 00905364. doi: 10.1214/aos/1176342360. URL http://projecteuclid.org/euclid.aos/1176342360. 4.1, 4.3

L A Foreman, A F M Smith, and I W Evett. Bayesian analysis of DNA profiling data in forensic identification applications. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 160(3):429–469, 1997.

R. C. Fuller. How and when selection experiments might actually be useful. *Integrative and Comparative Biology*, 45(3):391–404, June 2005. ISSN 15407063. doi: 10.1093/icb/45.3.391. URL http://icb.oxfordjournals.org/cgi/doi/10.1093/icb/45.3.391.

P Galbusera, L Lens, T Schenck, E Waiyaki, and E Matthysen. Genetic variability and gene flow

in the globally, critically-endangered Taita thrush. *Conservation Genetics*, 1:45–55, 2000. 4.5.2

Hong Gao, Katarzyna Bryc, and Carlos D Bustamante. On identifying the optimal number of population clusters via the deviance information criterion. *PloS one*, 6(6):e21014, January 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0021014. URL http://dx.plos.org/10.1371/journal.pone.0021014. 4.1

T Garland Jr and T Garland. Selection experiments: an under-utilized tool in biomechanics and organismal biology. In *Vertebrate Biomechanics and Evolution*, chapter 3, pages 23–57. BIOS Scientific Publishers, 2003. 6

R A Gibbs. The International HapMap Project. *Nature*, 426(6968):789–796, 2003. ISSN 14764687. doi: 10.1038/nature02168nature02168. 1, 5.1

D B Goldstein, A R Linares, L L Cavalli-Sforza, and M W Feldman. An evaluation of genetic distances for use with microsatellite loci. *Genetics*, 139(1):463–471, 1995.

R.E. Green, J. Krause, A.W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M.H.Y. Fritz, et al. A draft sequence of the neandertal genome. *science*, 328(5979): 710–722, 2010. 7

M F Hammer, T Karafet, A Rasanayagam, E T Wood, T K Altheide, T Jenkins, R C Griffiths, A R Templeton, and S L Zegura. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Molecular Biology and Evolution*, 15(4):427–441, 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/9549093. 1, 2.2, 3.5.2

James J Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979. URL http://www.jstor.org/stable/1912352. 5.2.2

S T Henderson and T D Petes. Instability of simple sequence DNA in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 12(6):2749–2757, 1992. 3.2.3

W G Hill and a Caballero. Artificial selection experiments. *Annual Review of Ecology and*

Systematics, 23(1):287–310, November 1992. ISSN 00664162. doi: 10.1146/annurev.es.23.110192.001443. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.es.23.110192.001443. 6

Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 106(23):9362–9367, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19474294. 1

Clive J Hoggart, Eteban J Parra, Mark D Shriver, Carolina Bonilla, Rick A Kittles, David G Clayton, and Paul M McKeigue. Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics*, 72(6):1492–1504, June 2003. ISSN 0002-9297. doi: 10.1086/375613. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1180309&tool=pmcentrez&rendertype=abstract. 5.2.2

R R Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44, 1990. 3.5.1

Richard R Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/18/2/337. 4.5.1

John P Huelsenbeck and Peter Andolfatto. Inference of population structure under a Dirichlet process prior. *Genetics*, 175(April):1787–1802, April 2007. ISSN 0016-6731. doi: 10.1534/genetics.106.061317. URL http://www.genetics.org/cgi/content/abstract/genetics.106.061317v1. 3.1, 3.6, 4.1, 4.4, 4.5.1, 4.5.2

Michael I Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999. ISSN

08856125. doi: 10.1023/A:1007665907178. URL http://www.springerlink.com/index/N811M25287935571.pdf. 3.3

R Kaeuffer, D Réale, D W Coltman, and D Pontier. Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity*, 99(4): 374–80, October 2007. ISSN 0018-067X. doi: 10.1038/sj.hdy.6801010. URL http://dx.doi.org/10.1038/sj.hdy.6801010. 5.2.1

Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18385116. 2.4, 6

R Kelly, M Gibbs, A Collick, and A J Jeffreys. Spontaneous mutation at the hypervariable mouse minisatellite locus Ms6-hm: flanking DNA sequence and analysis of germline and early somatic mutation events. *Proceedings: Biological Sciences*, 245(1314):235–245, 1991. 1, 3.2.3

A.B. Lee, D. Luca, L. Klei, B. Devlin, and K. Roeder. Discovering genetic ancestry using spectral graph theory. *Genetic Epidemiology*, 34(1):51–59, 2010.

Seunghak Lee and Eric P Xing. Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs. *Bioinformatics*, 28(12):i137–i146, June 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts227. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/28/12/i137. 6.2.1

Liming Liang, Sebastian Zöllner, and Gonçalo R Abecasis. GENOME: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–7, June 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm138. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/12/1565. 5.4.1

P Liang and Michael I Jordan. An asymptotic analysis of generative, discriminative, and pseu-

dolikelihood estimators. In *Proceedings of the 25th international conference on Machine learning*, pages 584–591. ACM, 2008. URL http://portal.acm.org/citation.cfm?id=1390230.

Cécile Libioulle, Edouard Louis, Sarah Hansoul, Cynthia Sandor, Frédéric Farnir, Denis Franchimont, Séverine Vermeire, Olivier Dewit, Martine de Vos, Anna Dixon, Bruno Demarche, Ivo Gut, Simon Heath, Mario Foglio, Liming Liang, Debby Laukens, Myriam Mni, Diana Zelenika, André Van Gossum, Paul Rutgeerts, Jacques Belaiche, Mark Lathrop, and Michel Georges. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genetics*, 3(4):e58, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17447842. 1

T Lin, E W Myers, and E P Xing. Interpreting anonymous DNA samples from mass disasters–probabilistic forensic inference using genetic markers. *Bioinformatics*, 22(14):e298, 2006. 3.2.3

Anthony D. Long and Charles H. Langley. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.*, 9(8):720–731, August 1999. doi: 10.1101/gr.9.8.720. URL http://genome.cshlp.org/cgi/content/abstract/9/8/720.

Trudy F C Mackay, Stephen Richards, Eric A Stone, Antonio Barbadilla, Julien F Ayroles, Dianhui Zhu, Sònia Casillas, Yi Han, Michael M Magwire, Julie M Cridland, Mark F Richardson, Robert R H Anholt, Maite Barrón, Crystal Bess, Kerstin Petra Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura Duncan, Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N Jhangiani, Katherine W Jordan, Fremiet Lara, Faye Lawrence, Sandra L Lee, Pablo Librado, Raquel S Linheiro, Richard F Lyman, Aaron J Mackey, Mala Munidasa, Donna Marie Muzny, Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel Ràmia, Jeffrey G Reid, Stephanie M Rollmann, Julio Rozas, Nehad Saada, Lavanya Turlapati, Kim C Worley, Yuan-Qing Wu, Akihiko Ya-

mamoto, Yiming Zhu, Casey M Bergman, Kevin R Thornton, David Mittelman, and Richard A Gibbs. The Drosophila melanogaster genetic reference panel. *Nature*, 482(7384):173–178, 2012. ISSN 00280836. doi: 10.1038/nature10811. URL http://www.nature.com/doifinder/10.1038/nature10811. 6.1

Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461 (7265):747–53, October 2009. ISSN 1476-4687. doi: 10.1038/nature08494. URL http://dx.doi.org/10.1038/nature08494. 1

N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209, 1967. 4.5.2, 5.5

Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John P A Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5):356–69, May 2008. ISSN 1471-0064. doi: 10.1038/nrg2344. URL http://dx.doi.org/10.1038/nrg2344.

Gil McVean. A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5 (10):e1000686, 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000686. URL http://dx.plos.org/10.1371/journal.pgen.1000686. 5.2.1, 5.2.2, 5.6

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):25, 2008. URL http://arxiv.org/abs/0811.2177. 6.2, 7

T P Minka. Estimating a Dirichlet distribution. 2000. 3.4.3

T. J. Morgan, T Garland, B L Irwin, J G Swallow, and P A Carter. The mode of evolution of molecular markers in populations of house mice under artificial selection for locomotor behavior. *The Journal of heredity*, 94(3):236–242, 2003. ISSN 1471-8505. doi: 10.1093/jhered/esg021. URL http://jhered.oupjournals.org/cgi/doi/10.1093/jhered/esg021.

John Novembre, Toby Johnson, Katarzyna Bryc, Zoltan Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008. ISSN 14764687. doi: 10.1038/nature07331. URL http://www.ncbi.nlm.nih.gov/pubmed/18758442. 1, 4.6

Ju-Hyun Park, Sholom Wacholder, Mitchell H Gail, Ulrike Peters, Kevin B Jacobs, Stephen J Chanock, and Nilanjan Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, 42(7):570–575, 2010. ISSN 15461718. doi: 10.1038/ng.610. URL http://www.ncbi.nlm.nih.gov/pubmed/20562874. 2.4.1

N Patterson, A L Price, and D Reich. Population structure and eigenanalysis. *PLoS Genet*, 2 (12):e190, 2006. 2.2, 3.1

J Pella and M Masuda. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(3):576–596, 2006. 3.1, 4.1

Steven J Phillips, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19323182. 5.2.2

D Pisani, L L Poling, M Lyons-Weiler, and S B Hedges. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biology*, 2004. 3.2.3

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006. ISSN 1061-4036. doi: 10.1038/ng1847. URL http://www.ncbi.nlm.nih.gov/pubmed/16862161. 2.4, 2.4.1, 6

J K Pritchard. Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69(1):124–137, 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11404818.

J K Pritchard, M Stephens, and P Donnelly. Inference of population structure from multilocus genotype data. *Genetics*, 155:945–959, 2000a. 3.3, 3.5.1, 4.5.2, 5.2.1, 5.2.2, 5.5

J K Pritchard, M Stephens, N A Rosenberg, and P Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000b. URL http://www.ncbi.nlm.nih.gov/pubmed/10827107. 2.2, 2.3, 2.4.1, 3, 3.1

K Puniyani, S Kim, and E P Xing. Multi-population GWA mapping via multi-task regularized regression. *Bioinformatics*, 26(12):i208–i216, 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btq191. URL http://www.bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq191. 2.4.1

D C Queller, J E Strassmann, and C R Hughes. Microsatellites and kinship. *Trends in Ecology & Evolution*, 8(8):285–288, 1993. 2.2, 3.2.3

Sohini Ramachandran, Omkar Deshpande, Charles C Roseman, Noah A Rosenberg, Marcus W Feldman, and L Luca Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15942–7, 2005. ISSN 00278424. doi: 10.1073/pnas.0507611102. URL http://www.ncbi.nlm.nih.

gov/pubmed/16243969. 4.6, 5.2.2

M. Rasmussen, X. Guo, Y. Wang, K.E. Lohmueller, S. Rasmussen, A. Albrechtsen, L. Skotte, S. Lindgreen, M. Metspalu, T. Jombart, et al. An aboriginal australian genome reveals separate human dispersals into asia. *Science*, 334(6052):94–98, 2011. 7

D. Reich, R.E. Green, M. Kircher, J. Krause, N. Patterson, E.Y. Durand, B. Viola, A.W. Briggs, U. Stenzel, P.L.F. Johnson, et al. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053–1060, 2010. 7

David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L Price, and Lalji Singh. Reconstructing Indian population history. *Nature*, 461(7263):489–94, 2009. ISSN 1476-4687. doi: 10.1038/nature08365. URL http://www.ncbi.nlm.nih.gov/pubmed/19779445. 1, 5.5.4

A Robertson. A theory of limits in artificial selection. 153:234–249, 1960.

K Roeder, M Escoar, J B Kadane, and I Balazs. Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika*, 85(2):269, 1998. 2.2, 2.4.1

N A Rosenberg, J K Pritchard, J L Weber, H M Cann, K K Kidd, L A Zhivotovsky, and M W Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002. 1, 2.2, 2.3, 2.3.2, 3, 3.1, 3.5.2, 4.5.2

Richa Saxena, Benjamin F Voight, Valeriya Lyssenko, Noël P Burtt, Paul I W de Bakker, Hong Chen, Jeffrey J Roix, Sekar Kathiresan, Joel N Hirschhorn, Mark J Daly, Thomas E Hughes, Leif Groop, David Altshuler, Peter Almgren, Jose C Florez, Joanne Meyer, Kristin Ardlie, Kristina Bengtsson Boström, Bo Isomaa, Guillaume Lettre, Ulf Lindblad, Helen N Lyon, Olle Melander, Christopher Newton-Cheh, Peter Nilsson, Marju Orho-Melander, Lennart Rå stam, Elizabeth K Speliotes, Marja-Riitta Taskinen, Tiinamaija Tuomi, Candace Guiducci, Anna Berglund, Joyce Carlson, Lauren Gianniny, Rachel Hackett, Liselotte Hall, Johan Holmkvist, Esa Laurila, Marketa Sjögren, Maria Sterner, Aarti Surti, Margareta Svensson, Malin Svens-

son, Ryan Tewhey, Brendan Blumenstiel, Melissa Parkin, Matthew Defelice, Rachel Barry, Wendy Brodeur, Jody Camarata, Nancy Chia, Mary Fava, John Gibbons, Bob Handsaker, Claire Healy, Kieu Nguyen, Casey Gates, Carrie Sougnez, Diane Gage, Marcia Nizzari, Stacey B Gabriel, Gung-Wei Chirn, Qicheng Ma, Hemang Parikh, Delwood Richardson, Darrell Ricke, and Shaun Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science New York NY*, 316(5829):1331–1336, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17463246. 1

G Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978. 3.2.3, 3.5.1, 4.1

Suyash Shringarpure and Eric P Xing. mStruct: Inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, 182(2):575–593, 2009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2691765&tool=pmcentrez&rendertype=abstracthttp://www.ncbi.nlm.nih.gov/pubmed/19363128. 1, 3, 3.4.4, 4.6, 7

Suyash Shringarpure and Eric P Xing. Artificial selection experiments for association in model organisms. Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15215, USA., 2012. 6.2

Suyash Shringarpure, Daegun Won, and Eric P Xing. StructHDP: automatic inference of number of clusters and population structure from admixed genotype data. *Bioinformatics*, 27(13):i324–32, July 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr242. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/27/13/i324. 1, 4, 7

Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J Hudson, Alexandre Montpetit, Alexey V Pshezhetsky, Marc Prentki, Barry I Posner, David J Balding, David Meyre, Constantin Polychronakos,

and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007. URL http://www.ncbi.nlm.nih.gov/pubmed/17293876. 1

K A Sohn and E P Xing. Spectrum: joint bayesian inference of population structure and recombination events. *Bioinformatics*, 23(13):i479—-i489, 2007. 3.6

Dennis Stanton and Dennis White. *Constructive combinatorics*. Undergraduate texts in mathematics. Springer-Verlag, 1986. 4.4.2

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445, 2003. URL http://www.ncbi.nlm.nih.gov/pubmed/12883005. 6

Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15712363. 1, 5.2.2, 5.5

Yee Whye Teh, Michael Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Science*, pages 1–41, 2005. 4.1, 4.3, 4.4.1, 4.4.1, 4.4.1, 4.4.1, 4.6

Yee Whye Teh, Kenichi Kurihara, and Max Welling. Collapsed variational inference for HDP. *Advances in Neural Information Processing Systems 20*, 20:1481–1488, 2008. URL http://eprints.pascal-network.org/archive/00003791/. 4.6

Alan Templeton. Out of Africa again and again. *Nature*, 416(6876):45–51, 2002. URL http://www.ncbi.nlm.nih.gov/pubmed/17708680. 1, 2.2, 3.5.2

Gilles Thomas, Kevin B Jacobs, Meredith Yeager, Peter Kraft, Sholom Wacholder, Nick Orr, Kai Yu, Nilanjan Chatterjee, Robert Welch, Amy Hutchinson, Andrew Crenshaw, Geraldine Cancel-Tassin, Brian J Staats, Zhaoming Wang, Jesus Gonzalez-Bosquet, Jun Fang, Xiang Deng, Sonja I Berndt, Eugenia E Calle, Heather Spencer Feigelson, Michael J Thun, Carmen Rodriguez, Demetrius Albanes, Jarmo Virtamo, Stephanie Weinstein, Fredrick R Schu-

macher, Edward Giovannucci, Walter C Willett, Olivier Cussenot, Antoine Valeri, Gerald L Andriole, E David Crawford, Margaret Tucker, Daniela S Gerhard, Joseph F Fraumeni, Robert Hoover, Richard B Hayes, David J Hunter, and Stephen J Chanock. Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, 40(3):310–315, 2008. ISSN 15461718. doi: 10.1038/ng.91. URL http://www.nature.com/ng/journal/v40/n3/abs/ng.91.html. 1

A M Valdes, M Slatkin, and N B Freimer. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics*, 133(3):737–749, 1993. 1, 3.2.3

Francis Vella. Estimating models with sample selection bias: a survey. *Journal of Human Resources*, 33(1):127–169, 1998. ISSN 0022166X. doi: 10.2307/146317. URL http://www.jstor.org/stable/146317?origin=crossref. 5.2.2

X Wan, Can Yang, Q Yang, H Xue, X Fan, N L S Tang, and W Yu. BOOST : A fast approach to detecting gene-gene interactions in. *Methods*, 87(3):1–4, 2010. URL http://arxiv.org/pdf/1001.5130. 6.2.1

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *Annals of Statistics*, 37(5A):2178–2201, 2007. URL http://arxiv.org/abs/0704.1139. 6.2, 7

M Y Wong, N E Day, J A Luan, and N J Wareham. Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Statistics in Medicine*, 23(6):987–998, 2004. 2.4

Eric P Xing, Michael Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of UAI*, pages 583–591, 2003. 3.3.1

Jian Yang, Beben Benyamin, Brian P Mcevoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(June):2010–2010, 2010. ISSN 10614036. doi:

10.1038/ng.608. URL http://www.nature.com/doifinder/10.1038/ng.608. 2.4.1, 7

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. *Twenty-first international conference on Machine learning ICML 04*, page 114, 2004. doi: 10.1145/1015330.1015425. URL http://portal.acm.org/citation.cfm?doid=1015330.1015425. 5.2.2

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining 2003 ICDM 2003 Third IEEE International Conference on*, volume 2003, pages 435–442. IEEE, 2003. ISBN 0769519784. doi: 10.1109/ICDM.2003.1250950. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1250950. 5.3.3

L A Zhivotovsky, P A Underhill, C Cinnioglu, M Kayser, B Morar, T Kivisild, R Scozzari, F Cruciani, G Destro-bisol, G Spedini, and Others. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *American Journal of Human Genetics*, 74(1):50–61, 2004. 1, 3.2.3

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA15213

Carnegie Mellon